

CONVERGENCE RATE ANALYSIS OF PRIMAL-DUAL SPLITTING SCHEMES*

DAMEK DAVIS†

Abstract. Primal-dual splitting schemes are a class of powerful algorithms that solve complicated monotone inclusions and convex optimization problems that are built from many simpler pieces. They decompose problems that are built from sums, linear compositions, and infimal convolutions of simple functions so that each simple term is processed individually via proximal mappings, gradient mappings, and multiplications by the linear maps. This leads to easily implementable and highly parallelizable or distributed algorithms, which often obtain nearly state-of-the-art performance.

In this paper, we analyze a monotone inclusion problem that captures a large class of primal-dual splittings as a special case. We introduce a unifying scheme and use some abstract analysis of the algorithm to prove convergence rates of the proximal point algorithm, forward-backward splitting, Peaceman-Rachford splitting, and forward-backward-forward splitting applied to the model problem. Our ergodic convergence rates are deduced under variable metrics, stepsizes, and relaxation. Our nonergodic convergence rates are the first shown in the literature. Finally, we apply our results to a large class of primal-dual algorithms that are a special case of our scheme and deduce their convergence rates.

Key words. primal-dual algorithms, convergence rates, proximal point algorithm, forward-backward splitting, forward-backward-forward splitting, Douglas-Rachford splitting, Peaceman-Rachford splitting, nonexpansive operator, averaged operator, fixed-point algorithm

AMS subject classifications. 47H05, 65K05, 65K15, 90C25

1. Introduction. Primal-dual algorithms are abstract splitting schemes that solve monotone inclusion and convex optimization problems. These schemes fully decompose problems built from sums, linear compositions, parallel sums, and infimal convolutions of simple functions so that each simple term is processed individually. This decomposition is achieved by cleverly combining primal and dual pair problems into a single inclusion problem, to which standard operator splitting algorithms can be applied. This process gives rise to algorithms that are inherently parallel or distributed and in which expensive matrix inversions can be avoided. The characteristics of primal-dual algorithms are especially desirable for large-scale applications in machine learning, image processing, distributed optimization, and control.

Primal-dual methods have a long history with many contributors, and an attempt to summarize and relate all of the contributions is beyond the scope of this paper. In this paper, we are mainly concerned with the line of work that began in [41, 15, 25] and the many generalizations and enhancements of the basic framework that followed [19, 22, 46, 12, 9, 10, 17, 31, 6, 18]. Thus, we consider the following prototypical convex optimization problem as our guiding example:

$$\underset{x \in \mathcal{H}_0}{\text{minimize}} \ f(x) + g(x) + \sum_{i=1}^n (h_i \square l_i)(B_i x) \quad (1.1)$$

where \square denotes the infimal convolution operation (see Section 1.2), $n \in \mathbf{N}$, $n \geq 1$, \mathcal{H}_i are Hilbert spaces for $i = 0, \dots, n$, the functions $f, g : \mathcal{H}_0 \rightarrow (-\infty, \infty]$ and $h_i, l_i :$

*This work is partially supported by NSF GRFP grant DGE-0707424.

†Department of Mathematics, University of California, Los Angeles, Los Angeles, CA 90025/
School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850
(damek@math.ucla.edu)

$\mathcal{H}_i \rightarrow (-\infty, \infty]$ are closed, proper, and convex for $i = 1, \dots, n$, and $B_i : \mathcal{H}_0 \rightarrow \mathcal{H}_i$ is a bounded linear map for $i = 1, \dots, n$.

All of the algorithms presented in this paper completely disentangle the structure of Problem (1.1) so that each iteration only involves the individual proximal operators of each of the nondifferentiable terms, the gradient operators of the differentiable terms, and multiplication by the linear maps. Thus, the maps B_i are never inverted, and we never compute proximal operators or gradients of sums or infimal convolutions of functions. We note that this level of separability is not achieved by classical splitting methods such as forward-backward splitting, Douglas-Rachford splitting, or the alternating direction method of multipliers (ADMM) when they are applied directly to the primal optimization Problem (1.1) [13, 38, 26, 33].

In Problem (1.1), the maps B_i can be used as “data matrices,” in which case h_i and l_i are *data fitting* terms and f and g enforce *prior knowledge* on the structure of the solution, such as sparsity, low rank, or smoothness. In other cases, the maps h_i and l_i may be regularizers that emphasize many competing structures. We now present an example.

Application: Constrained model fitting with group-structured regularizers. Fix $d, m \in \mathbf{N} \setminus \{0\}$. Suppose we are given a measurement $b \in \mathbf{R}^d$ and a dictionary $A \in \mathbf{R}^{d \times m}$. Our goal is to recover a highly structured signal $x = (x_1, \dots, x_m)^T \in \mathbf{R}^m$ such that $Ax \approx b$. For example, in the hierarchical sparse coding problem (HSCP) [29], we arrange the columns of A into a directed tree structure \mathcal{T} and allow $x_i = 0$ only if $x_j = 0$ for all descendants j in \mathcal{T} of node i . Such a hierarchical representation is particularly useful for multi-scale data such as images and text documents. This type of regularization can be generalized to include arbitrary column groupings and complicated relationships between the elements of each group. Indeed, let G be a set of (possibly overlapping) subsets of $\{1, \dots, m\}$. For all $S \in G$ and $x \in \mathbf{R}^m$, let $B_S x = L_S(x)_{i \in S}^T \in \mathbf{R}^{m_S}$ where $m_S \in \mathbf{N} \setminus \{0\}$ and $L_S : \mathbf{R}^{|S|} \rightarrow \mathbf{R}^{m_S}$ is a linear map. Let $C \subseteq \mathbf{R}^m$ be a closed convex set, and let $\iota_C : \mathbf{R}^m \rightarrow \{0, \infty\}$ be the convex indicator function of C . For all $S \in G$, let $h_S : \mathbf{R}^{m_S} \rightarrow (-\infty, \infty]$ be a closed, proper, and convex regularizer, and let $l_S = \iota_{\{0\}}$, which implies $h_S \square l_S = h_S$. Then one special case of Problem (1.1) is the group-structured regularized model fitting problem:

$$\underset{x \in \mathbf{R}^m}{\text{minimize}} \quad \iota_C(x) + (1/2)\|Ax - b\|^2 + \sum_{S \in G} h_S(B_S x).$$

In [29], the authors consider the nonnegativity constraint $C = \mathbf{R}_{\geq 0}^m$ and a grouping G which consists of overlapping sets S_i for $i \in \{1, \dots, m\}$ such that S_i contains i and all of the descendants of i in \mathcal{T} . Furthermore, for each $S \in G$, they consider the map $L_S = I_{\mathbf{R}^{|S|}}$ and the function $h_S = w_S \|(x_i)_{i \in S}^T\|_p$ where $p \in [1, \infty]$ and $w_S > 0$. This setup induces a mixed ℓ_1/ℓ_p norm on \mathbf{R}^m of the form $\sum_{S \in G} w_S \|(x_i)_{i \in S}^T\|_p$, which tends to “zero out” entire groups of components. Note that the sum is also highly nonseparable in the components of x , which can make the proximal operator of the regularization term difficult to evaluate. If we denote $f(x) = \iota_C(x)$ and $g(x) = (1/2)\|Ax - b\|^2$, then the algorithms in this paper only utilize the projection $P_C = \mathbf{prox}_f$ onto C , the gradient $\nabla g(x) = A^*(Ax - b)$, and for all $S \in G$ in parallel, multiplications by the maps B_S and B_S^* , and evaluations of the proximal operator of the function h_S . Not only does this make each iteration of the algorithm simple to implement and computationally inexpensive, it also provides a unified algorithmic framework for higher order regularizations of the components in each group, a task which might otherwise be intractable in large-scale applications.

Finally, we note that the use of infimal convolutions in applications is not widespread, so we list a few instances where they may be useful: Infimal convolutions are used in image recovery [14, Section 5] to remove staircasing effects in the total variation model. The infimal convolution of the indicator functions of two closed convex sets is the indicator function of their Minkowski sum, which has applications in motion planning for robotics [32, Section 4.3.2]. In convex analysis, the Moreau envelope of a function arises as an infimal convolution with a multiple of the squared norm [2, Section 12.4]. More generally, the infimal convolution of h_i and l_i can be interpreted as a regularization or smoothing of h_i by l_i and vice versa [2, Section 18.3].

1.1. Goals, challenges, and approaches. This work seeks to improve the theoretical understanding of the convergence rates of primal-dual splitting schemes. In this paper, we study primal-dual algorithms that are applications of standard operator splitting algorithms in product spaces consisting of primal and dual variables. Consequently, the convergence theory for these algorithms is well-developed, and they are known to converge (weakly) under mild conditions.

Although we understand when these algorithms converge, relatively little is known about their rate of convergence. For convex optimization algorithms, the *ergodic* convergence rate of the *primal-dual gap* has been analyzed in a few cases [15, 7, 6, 43]. However, even in cases where convergence rates are known, variable metrics and stepsizes, which can significantly improve practical performance of the algorithms [40, 27], are not analyzed. In addition, we are not aware of any convergence rate analysis of the primal-dual gap for the *nonergodic* (or last) iterate generated by these algorithms. It is important to understand nonergodic convergence rates because the ergodic (or time-averaged) iterates can “average out” structural properties, such as sparsity and low rank, that are shared by the solution and the nonergodic iterate.

The convergence rate analysis of the ergodic primal-dual gap largely follows from subgradient inequalities and an application of Jensen’s inequality. In contrast, the techniques developed in this paper exploit the properties of the nonexpansive operators driving the algorithms to deduce the nonergodic convergence rate of the primal-dual gap. Thus, our techniques are quite different from those used in classical convergence rate analysis and parallel the analysis developed in [24].

We summarize our contributions and techniques as follows:

(i) We describe a model monotone inclusion problem that generalizes many primal-dual formulations that appear in the literature. We provide a simple prototype algorithm to solve the model problem, and we deduce a fundamental inequality that bounds the primal-dual gap at each iteration of the algorithm. We then simplify the inequality in the special case of four splitting algorithms (Section 2).

(ii) We derive ergodic convergence rates of the variable metric forms of the relaxed proximal point algorithm (PPA), relaxed forward-backward splitting (FBS), and forward-backward-forward splitting as well as the fixed metric relaxed Peaceman-Rachford splitting (PRS) algorithm (Section 3). After some algebraic simplifications, our analysis essentially follows from an application of Jensen’s inequality.

(iii) We derive nonergodic convergence rates of relaxed PPA, relaxed FBS, and relaxed PRS (Section 4). All of our analysis follows by bounding the primal-dual gap function by a multiple of the *fixed-point residual* (FPR) of the nonexpansive mapping that drives the algorithm. Thus, we show that the size of the FPR can be used as a valid *stopping criteria* for these three algorithms.

(iv) We apply our results to deduce ergodic and nonergodic convergence rates for a large class of primal-dual algorithms that have appeared in the literature (Section 5).

Our analysis not only deduces the convergence rates of a large class of primal-dual algorithms found in the literature. It also serves as a resource for the analysis of future primal-dual algorithms that solve generalizations of Problem 1.1, e.g., [3, 10].

1.2. Definitions, notation and some facts. In what follows, \mathcal{H}, \mathcal{G} , and \mathbf{H} denote (possibly infinite dimensional) Hilbert spaces. We always use the notations $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ to denote the inner product and norm associated to a Hilbert space, respectively. Note that there is some ambiguity in this convention, but it simplifies the notation and no confusion should arise. The space \mathbf{H} will usually denote a product Hilbert space consisting of primal variables in \mathcal{H} and dual variables in \mathcal{G} . Let $\mathbf{R}_{++} = \{x \in \mathbf{R} \mid x > 0\}$ denote the set of strictly positive real numbers. Let $\mathbf{N} = \{k \in \mathbf{Z} \mid k \geq 0\}$ denote the set of nonnegative integers. In all of the algorithms we consider, we utilize two stepsize sequences: the implicit sequence $(\gamma_j)_{j \in \mathbf{N}} \subseteq \mathbf{R}_{++}$ and the explicit sequence $(\lambda_j)_{j \in \mathbf{N}} \subseteq \mathbf{R}_{++}$. We define the k -th partial sum of the sequence $(\gamma_j \lambda_j)_{j \in \mathbf{N}}$ by the formula:

$$\Sigma_k := \sum_{i=0}^k \gamma_i \lambda_i. \quad (1.2)$$

Given a sequence $(x^j)_{j \in \mathbf{N}} \subset \mathcal{H}$ and $k \in \mathbf{N}$, we let $\bar{x}^k = (1/\Sigma_k) \sum_{i=0}^k \gamma_i \lambda_i x^i$ denote its k th average with respect to the sequence $(\gamma_j \lambda_j)_{j \in \mathbf{N}}$. We call a convergence result *ergodic* if it is in terms of the sequence $(\bar{x}^j)_{j \in \mathbf{N}}$, and *nonergodic* if it is in terms of $(x^j)_{j \in \mathbf{N}}$.

We denote the set of summable nonnegative sequences by $\ell_+^1(\mathbf{N}) := \{(\eta_j)_{j \in \mathbf{N}} \subseteq [0, \infty) \mid \sum_{j=0}^{\infty} \eta_j < \infty\}$.

The following definitions and facts are mostly standard and can be found in [2, 20]

We let $\mathcal{B}(\mathcal{H}, \mathcal{G})$ denote the set of bounded linear maps from \mathcal{H} to \mathcal{G} , and set $\mathcal{B}(\mathcal{H}) := \mathcal{B}(\mathcal{H}, \mathcal{H})$. We will use the notation $I_{\mathcal{H}} \in \mathcal{B}(\mathcal{H})$ to denote the identity map. Given a map $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$, we denote its adjoint by $L^* \in \mathcal{B}(\mathcal{G}, \mathcal{H})$. The operator norm on $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$ is defined by the following supremum: $\|L\| = \sup_{x \in \mathcal{H}, \|x\| \leq 1} \|Lx\|$. Let $\rho \in \mathbf{R}_+$ be a nonnegative real number. We let $\mathcal{S}_{\rho}(\mathcal{H}) \subseteq \mathcal{B}(\mathcal{H})$ denote the set of linear ρ -strongly monotone self-adjoint maps:

$$\mathcal{S}_{\rho}(\mathcal{H}) := \{U \in \mathcal{B}(\mathcal{H}) \mid U = U^*, (\forall x \in \mathcal{H}) \langle Ux, x \rangle \geq \rho \|x\|^2\}.$$

We define the (semi)-norm and inner product induced by $U \in \mathcal{S}_{\rho}(\mathcal{H})$ on \mathcal{H} by the formulae: for all $x, y \in \mathcal{H}$, $\|x\|_U^2 := \langle Ux, x \rangle$, and $\langle x, y \rangle_U := \langle Ux, y \rangle$. The Loewner partial ordering on $\mathcal{S}_{\rho}(\mathcal{H})$ is defined as follows: for all $U_1, U_2 \in \mathcal{S}_{\rho}(\mathcal{H})$, we have

$$U_1 \succcurlyeq U_2 \quad \Longleftrightarrow \quad (\forall x \in \mathcal{H}) \quad \|x\|_{U_1}^2 \geq \|x\|_{U_2}^2.$$

Let $L \geq 0$, and let D be a nonempty subset of \mathcal{H} . A map $T : D \rightarrow \mathcal{H}$ is called L -Lipschitz if for all $x, y \in D$, we have $\|Tx - Ty\| \leq L\|x - y\|$. In particular, T is called *nonexpansive* if it is 1-Lipschitz. A map $N : D \rightarrow \mathcal{H}$ is called λ -averaged [2, Section 4.4] if there exists a nonexpansive map $T : D \rightarrow \mathcal{H}$ and $\lambda \in (0, 1)$ such that

$$N = T_{\lambda} := (1 - \lambda)I_{\mathcal{H}} + \lambda T. \quad (1.3)$$

A $(1/2)$ -averaged map is called *firmly nonexpansive*.

Let $2^{\mathcal{H}}$ denote the power set of \mathcal{H} . A set-valued operator $A : \mathcal{H} \rightarrow 2^{\mathcal{H}}$ is called *monotone* if for all $x, y \in \mathcal{H}$, $u \in Ax$, and $v \in Ay$, we have $\langle x - y, u - v \rangle \geq 0$. We

denote the set of zeros of a monotone operator by $\text{zer}(A) := \{x \in \mathcal{H} \mid 0 \in Ax\}$. The *graph* of A is denoted by $\text{gra}(A) := \{(x, y) \mid x \in \mathcal{H}, y \in Ax\}$. Evidently, A is uniquely determined by its graph. A monotone operator A is called *maximal monotone* provided that $\text{gra}(A)$ is not properly contained in the graph of any other monotone set-valued operator. The *inverse* of A , denoted by A^{-1} , is defined uniquely by its graph: $\text{gra}(A^{-1}) := \{(y, x) \mid x \in \mathcal{H}, y \in Ax\}$. Let $\beta \in \mathbf{R}_{++}$ be a positive real number. The operator A is called β -strongly monotone provided that for all $x, y \in \mathcal{H}$, $u \in Ax$, and $v \in Ay$, we have $\langle x - y, u - v \rangle \geq \beta \|x - y\|^2$. A *single-valued* operator $B : \mathcal{H} \rightarrow 2^{\mathcal{H}}$ maps each point in \mathcal{H} to a singleton and will be identified with the natural \mathcal{H} -valued map it defines. A single-valued operator B is called β -cocoercive provided that for all $x, y \in \mathcal{H}$, we have $\langle x - y, Bx - By \rangle \geq \beta \|Bx - By\|^2$. Evidently, B is β -cocoercive whenever B^{-1} is β -strongly monotone. The parallel sum of (not necessarily single-valued) monotone operators A and B is given by $A \square B := (A^{-1} + B^{-1})^{-1}$. The *resolvent* of a monotone operator A is defined by the inversion $J_A := (I + A)^{-1}$. Minty's theorem shows that J_A is single-valued and has full domain \mathcal{H} if, and only if, A is maximally monotone. Note that A is monotone if, and only if, J_A is firmly nonexpansive. Thus, the *reflection operator*

$$\text{refl}_A := 2J_A - I_{\mathcal{H}} \quad (1.4)$$

is nonexpansive on \mathcal{H} whenever A is maximally monotone. If $\rho > 0$ and $U \in \mathcal{S}_{\rho}(\mathcal{H})$, the operator $U^{-1}A$ is maximal monotone in $\langle \cdot, \cdot \rangle_U$, if, and only if, A is maximally monotone in $\langle \cdot, \cdot \rangle$. Let $\gamma \in (0, \infty)$. The resolvent of the map $\gamma U^{-1}A$ has the special identity: $J_{\gamma U^{-1}A} = U^{-1/2} J_{\gamma U^{-1/2} A U^{-1/2}} U^{1/2}$ [21, Example 3.9].

Let $\Gamma_0(\mathcal{H})$ denote the set of closed, proper, and convex functions $f : \mathcal{H} \rightarrow (-\infty, \infty]$. Let $\text{dom}(f) := \{x \in \mathcal{H} \mid f(x) < \infty\}$. We will let $\partial f(x) : \mathcal{H} \rightarrow 2^{\mathcal{H}}$ denote the subdifferential of f : $\partial f(x) := \{u \in \mathcal{H} \mid \forall y \in \mathcal{H}, f(y) \geq f(x) + \langle y - x, u \rangle\}$. We will always let

$$\tilde{\nabla} f(x) \in \partial f(x) \quad (1.5)$$

denote a subgradient of f drawn at the point x , and the actual choice of the subgradient $\tilde{\nabla} f(x)$ will always be clear from the context; note that this notation was also used in [4]. The subdifferential operator of f is maximally monotone. The inverse of ∂f is given by ∂f^* where $f^*(y) := \sup_{x \in \mathcal{H}} \{\langle y, x \rangle - f(x)\}$ is the *Fenchel conjugate* of f . If the function f is β -strongly convex, then ∂f is β -strongly monotone.

If a convex function $f : \mathcal{H} \rightarrow (-\infty, \infty]$ is Fréchet differentiable at $x \in \mathcal{H}$, then $\partial f(x) = \{\nabla f(x)\}$. Suppose f is convex and Fréchet differentiable on \mathcal{H} , and let $\beta \in \mathbf{R}_{++}$ be a positive real number. Then the Baillon-Haddad theorem states that ∇f is $(1/\beta)$ -Lipschitz, if, and only if, ∇f is β -cocoercive.

The resolvent operator associated to ∂f is called the *proximal operator* and is uniquely defined by the following (strongly convex) minimization problem: $\text{prox}_f(x) := J_{\partial f}(x) = \arg \min_{y \in \mathcal{H}} \{f(y) + (1/2)\|y - x\|^2\}$. If $\rho > 0$, $U \in \mathcal{S}_{\rho}(\mathcal{H})$, and $\gamma \in (0, \infty)$, the proximal operator of f in the metric induced by U is given by the following formula: for all $x \in \mathcal{H}$,

$$\text{prox}_{\gamma f}^U(x) := J_{\gamma U^{-1} \partial f}(x) = \arg \min_{y \in \mathcal{H}} \left\{ f(y) + \frac{1}{2\gamma} \|y - x\|_U^2 \right\}. \quad (1.6)$$

The *infimal convolution* of two functions $f, g : \mathcal{H} \rightarrow (-\infty, \infty]$ is denoted by $f \square g : \mathcal{H} \rightarrow [-\infty, \infty] : x \mapsto \inf_{y \in \mathcal{H}} \{f(y) + g(x - y)\}$. The indicator function of a closed,

convex set $C \subseteq \mathcal{H}$ is denoted by $\iota_C : \mathcal{H} \rightarrow \{0, \infty\}$; the indicator function is 0 on C and is ∞ on $\mathcal{H} \setminus C$.

We will always use a $*$ superscript to denote a fixed point of a nonexpansive map, a zero of a monotone inclusion, or a minimizer of an optimization problem, e.g., z^* .

Finally, we call the following identity the *cosine rule*:

$$(\forall x, y, z \in \mathcal{H}) \quad \|y - z\|^2 + 2\langle y - x, z - x \rangle = \|y - x\|^2 + \|z - x\|^2. \quad (1.7)$$

1.3. Assumptions.

ASSUMPTION 1 (Convexity). *Every function we consider is closed, proper, and convex.*

Unless otherwise stated, a function is not necessarily differentiable.

ASSUMPTION 2 (Differentiability). *Every differentiable function we consider is Fréchet differentiable [2, Definition 2.45].*

We employ other assumptions throughout the paper, but we list them closer to where they are invoked.

1.4. Basic properties of metrics. A simple proof of the following Lemma recently appeared in [20, Lemma 2.1]. It previously appeared in [30, Section VI.2.6].

LEMMA 1.1 (Metric properties). *Whenever $U, V \in \mathcal{S}_0(\mathcal{H})$ satisfy the inequality $\alpha I_{\mathcal{H}} \succcurlyeq U \succcurlyeq V \succcurlyeq \beta I_{\mathcal{H}}$ for $\alpha, \beta > 0$, we have the ordering $(1/\beta)I_{\mathcal{H}} \succcurlyeq V^{-1} \succcurlyeq U^{-1} \succcurlyeq (1/\alpha)I_{\mathcal{H}}$, the inclusion $U^{-1} \in \mathcal{S}_{\|U\|^{-1}}(\mathcal{H})$, and the inequality $\|U^{-1}\| \leq (1/\beta)$.*

1.5. Basic properties of resolvents and averaged operators.

The following are simple modifications of standard facts found in [2].

PROPOSITION 1.2. *Let $\rho > 0$, let $\lambda > 0$, let $\alpha \in (0, 1)$, let $U \in \mathcal{S}_\rho(\mathcal{H})$, let $A : \mathcal{H} \rightarrow \mathcal{H}$ be a single-valued maximal monotone operator, and let $f \in \Gamma_0(\mathcal{H})$*

1. Optimality conditions of J : *We have $x^+ := J_{\gamma U^{-1}(\partial f + A)}(x)$ if, and only if, there exists a unique subgradient $\tilde{\nabla} f(x^+) := (1/\gamma)U(x - x^+) - Ax^+ \in \partial f(x^+)$, such that*

$$\tilde{\nabla} f(x^+) + Ax^+ = \frac{1}{\gamma}U(x - x^+) \in \partial f(x^+) + Ax^+.$$

2. Averaged operator contraction property: *Let $\lambda \in (0, 1)$. A map $T : \mathcal{H} \rightarrow \mathcal{H}$ is λ -averaged in the metric induced by U if, and only if, for all $x, y \in \mathcal{H}$,*

$$\|Tx - Ty\|_U^2 \leq \|x - y\|_U^2 - \frac{1 - \lambda}{\lambda} \|(I_{\mathcal{H}} - T)x - (I_{\mathcal{H}} - T)y\|_U^2. \quad (1.8)$$

3. Wider relaxations: *A map $T : \mathcal{H} \rightarrow \mathcal{H}$ is α -averaged in $\|\cdot\|_U$, if, and only if, T_λ (Equation (1.3)) is $\lambda\alpha$ -averaged in $\|\cdot\|_U$ for all $\lambda \in (0, 1/\alpha)$. In addition, $T_{1/\alpha}$ is nonexpansive with respect to $\|\cdot\|_U$.*

1.6. Variable metrics. Throughout this paper we will consider sequences of mappings $(U_j)_{j \in \mathbb{N}} \in \mathcal{S}_\rho(\mathcal{H})$ for some $\rho > 0$. In order to apply the standard convergence theory for variable metrics, we will make the following assumption:

ASSUMPTION 3. *There exists a summable sequence $(\eta_j)_{j \in \mathbb{N}} \subseteq \ell_+^1(\mathbb{N})$ such that for all $k \in \mathbb{N}$, $(1 + \eta_k)U_k \succcurlyeq U_{k+1}$. In addition $\mu := \sup_{j \in \mathbb{N}} \|U_j\| < \infty$.*

Assumption 3 is standard in variable metric algorithms [20, 45, 21, 37].

REMARK 1. *There is an asymmetry in our notation and the notation of [20, 45, 21, 37]. In our analysis, the map $U \in \mathcal{S}_\rho(\mathcal{H})$ induces a metric on \mathcal{H} . In other papers, the maps U^{-1} induce a metric on \mathcal{H} .*

The following notation will be used throughout the rest of the paper. The proof is elementary.

PROPOSITION 1.3 (Metric parameters). *Suppose that $(\eta_j)_{j \in \mathbf{N}} \subseteq \ell_+^1(\mathbf{N})$. Define*

$$\eta_p := \prod_{i=0}^{\infty} (1 + \eta_i) \quad \text{and} \quad \eta_s := \sum_{i=0}^{\infty} \eta_i.$$

Then η_p and η_s are finite.

The following Proposition is a consequence of the proof of [20, Theorem 5.1]. The proof is simple, so we omit it.

PROPOSITION 1.4. *Let \mathcal{H} be a Hilbert space. Let $\rho \in (0, \infty)$, let $(\eta_j)_{j \in \mathbf{N}} \subseteq \ell_+^1(\mathbf{N})$, and let $(U_j)_{j \in \mathbf{N}} \in \mathcal{S}_\rho(\mathcal{H})$ satisfy Assumption 3. For all $k \in \mathbf{N}$, let $\alpha_k \in (0, 1)$, let $\lambda_k \in (0, 1/\alpha_k]$ be a relaxation parameter, and let $T_k : \mathcal{H} \rightarrow \mathcal{H}$ be α_k -averaged in the metric $\|\cdot\|_{U_k}$. Furthermore, assume that there is a point $z^* \in \mathcal{H}$ such that $T_k z^* = z^*$ for all $k \in \mathbf{N}$. Let the $(z^j)_{j \in \mathbf{N}}$ be generated by the following Krasnosel'skiĭ-Mann (KM)-type iteration (Equation (1.3)): let $z^0 \in \mathcal{H}$, and for all $k \in \mathbf{N}$, define*

$$z^{k+1} = (T_k)_{\lambda_k} z^k.$$

Then the following are true:

1. *For all $k \in \mathbf{N}$, $\|z^{k+1} - z^*\|_{U_{k+1}}^2 \leq (1 + \eta_k) \|z^k - z^*\|_{U_k}^2$ and hence,*

$$\|z^k - z^*\|_{U_k}^2 \leq \eta_p \|z^0 - z^*\|_{U_0}^2.$$

2. *The following sum is finite:*

$$\sum_{i=0}^{\infty} \frac{1 - \alpha_i \lambda_i}{\alpha_i \lambda_i} \|z^{i+1} - z^i\|^2 \leq \frac{1}{\rho} (1 + \eta_p \eta_s) \|z^0 - z^*\|_{U_0}^2.$$

We will use the following proposition to select parameters in the FBS algorithm. The proof of the following fact follows from [46, Equation (3.35)]

PROPOSITION 1.5. *Let $\rho > 0$, let $\beta > 0$, let $B : \mathcal{H} \rightarrow \mathcal{H}$ be β -cocoercive in the norm $\|\cdot\|$, and let $U \in \mathcal{S}_\rho(\mathcal{H})$. Then $U^{-1}B$ is $\beta\rho$ -cocoercive in the norm $\|\cdot\|_U$.*

The following proposition essentially follows from the proof of [45, Theorem 3.1].

PROPOSITION 1.6. *Let $A : \mathcal{H} \rightarrow 2^{\mathcal{H}}$ be maximal monotone, let $B : \mathcal{H} \rightarrow \mathcal{H}$ be monotone and $(1/\beta)$ -Lipschitz for some $\beta > 0$, let $\rho > 0$, let $(U_j)_{j \in \mathbf{N}} \subseteq \mathcal{S}_\rho(\mathcal{H})$ satisfy Assumption 3, and let $(\gamma_j)_{j \in \mathbf{N}} \subseteq (0, \rho\beta]$. Let $(z^j)_{j \in \mathbf{N}}$ be a sequence of points defined by the iteration: let $z^0 \in \mathcal{H}$ and for all $k \in \mathbf{N}$, define*

$$\begin{aligned} y^k &= z^k - \gamma_k U_k^{-1} B z^k; \\ x^k &= J_{\gamma_k U_k^{-1} A}(y^k); \\ w^k &= x^k - \gamma_k U_k^{-1} B x^k; \\ z^{k+1} &= z^k - y^k + w^k. \end{aligned}$$

Suppose that $\text{zer}(A + B) \neq \emptyset$. Then for all $z^ \in \text{zer}(A + B)$ and for all $k \in \mathbf{N}$, we have, $\|z^{k+1} - z^*\|_{U_{k+1}}^2 \leq (1 + \eta_k) \|z^k - z^*\|_{U_k}^2$.*

2. The unifying scheme. In this section, we introduce a prototype monotone inclusion problem that generalizes and summarizes many primal-dual problem formulations found in the literature. After we describe the problem, we will introduce an abstract unifying scheme that generalizes many existing primal-dual algorithms. We will describe how to measure convergence of the unifying scheme, and introduce a fundamental inequality that bounds our measure of convergence. Finally, we will identify the key terms in the fundamental inequality and simplify them in the case of several abstract splitting algorithms.

In Section 5, we will show that this unifying scheme relates to many existing algorithms, and extend the convergence rate results of those methods.

2.1. Problem and algorithm. We focus on the following problem:

PROBLEM 1 (Prototype primal-dual problem). *Let $(\mathbf{H}, \langle \cdot, \cdot \rangle)$ be a Hilbert space, let $\mathbf{f}, \mathbf{g} \in \Gamma_0(\mathbf{H})$, and let $\mathbf{S} : \mathbf{H} \rightarrow \mathbf{H}$ be a skew symmetric map: $\mathbf{S}^* = -\mathbf{S}$. Then the prototype primal-dual problem is to find $\mathbf{x}^* \in \mathbf{H}$ such that*

$$0 \in \partial \mathbf{f}(\mathbf{x}^*) + \partial \mathbf{g}(\mathbf{x}^*) + \mathbf{S} \mathbf{x}^*. \quad (2.1)$$

Evidently, Problem 1 is a *monotone inclusion problem* because $\partial \mathbf{f}$, $\partial \mathbf{g}$, and \mathbf{S} are maximally monotone operators on \mathbf{H} [2, Example 20.30].

We are now ready to define our unifying scheme.

Algorithm 1: Unifying scheme

input : $\mathbf{z}^0 \in \mathbf{H}; (\lambda_j)_{j \geq 0} \subseteq \mathbf{R}_{++}; (\gamma_j)_{j \in \mathbf{N}} \subseteq \mathbf{R}_{++}; \rho > 0; (U_j)_{j \in \mathbf{N}} \subseteq \mathcal{S}_\rho(\mathbf{H})$.
for $k = 0, 1, \dots$ **do**
 $\mathbf{z}^{k+1} = \mathbf{z}^k - \gamma_k \lambda_k U_k^{-1} \left(\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) + \tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) + \mathbf{S} \mathbf{x}_s^k \right);$

Note that the points $\mathbf{x}_f^k, \mathbf{x}_g^k$, and \mathbf{x}_s^k as well as the subgradients $\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) \in \partial \mathbf{f}(\mathbf{x}_f^k)$ and $\tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) \in \partial \mathbf{g}(\mathbf{x}_g^k)$ are unspecified in the description of Algorithm 1. In the algorithms we study, these points and subgradients will be generated by proximal and forward gradient operators and, thus, can be determined given \mathbf{z}^k ; see Section 2.2 for examples. However, Algorithm 1 is only meant to illustrate the algebraic form that our analysis addresses, and it is not meant to be an actual algorithm that solves Problem 2.1. The positive scalar sequence $(\lambda_j)_{j \in \mathbf{N}}$ consists of *relaxation parameters*, or explicit stepsize parameters, whereas the sequence $(\gamma_j)_{j \in \mathbf{N}}$ consists of *proximal parameters*, or implicit stepsize parameters. The strongly monotone maps $(U_j)_{j \in \mathbf{N}}$ induce the metrics used in each iteration of the algorithm.

In all of our applications, \mathbf{H} will be a product space of primal and dual variables. In this setting, \mathbf{f} and \mathbf{g} will be block-separable maps, and \mathbf{g} will sometimes be differentiable. The map \mathbf{S} “mixes” the primal and dual variable sequences in the product space. Mixing is necessary, because the sequences are otherwise uncoupled.

The sequence of maps $(U_j)_{j \in \mathbf{N}}$ is employed for two purposes. First, the maps are used because the evaluation of the resolvent $J_{\partial \mathbf{f} + \mathbf{S}}$, which is a basic building block of most of the algorithms we study, may not be simple. Thus, the primal-dual algorithms that we study formulate special metrics induced by $U \in \mathcal{S}_\rho(\mathbf{H})$ such that $J_{U^{-1}(\partial \mathbf{f} + \mathbf{S})}$ is as easy to evaluate as $\text{prox}_{\mathbf{f}}$ (See Section 5). Hence, in our analysis we must at least consider fixed metrics that are different from the standard product metric on \mathbf{H} . Second, we allow the metrics to vary at each iteration because it can significantly improve the practical performance of the algorithm, e.g., by employing second order information, or even simple time-varying diagonal metrics [40, 27].

2.2. Examples of the unifying scheme. In this section we introduce four algorithms and show that they are special cases of Algorithm 1. We will also introduce several assumptions on the algorithm parameters that ensure convergence. These assumptions will remain in effect throughout the rest of the paper. Note that the convergence theory of the methods in this section is well-studied. See [2, 20, 46, 21, 42, 44, 33] for background. Finally, we will say that several algorithms in this section are *relaxed*. For brevity, we will drop this adjective whenever convenient.

The relaxed variable metric PPA applies to problems in which $\mathbf{g} \equiv 0$.

Algorithm 2: Relaxed variable metric proximal point algorithm (PPA)

input : $\mathbf{z}^0 \in \mathbf{H}; (\lambda_j)_{j \geq 0} \subseteq (0, 2]; (\gamma_j)_{j \in \mathbf{N}} \subseteq \mathbf{R}_{++}; \rho > 0; (U_j)_{j \in \mathbf{N}} \subseteq \mathcal{S}_\rho(\mathbf{H})$.
for $k = 0, 1, \dots$ **do**
 $\mathbf{x}_g^k = \mathbf{z}^k$;
 $\mathbf{x}_f^k = J_{\gamma_k U_k^{-1}(\partial \mathbf{f} + \mathbf{S})}(\mathbf{z}^k)$;
 $\mathbf{z}^{k+1} = (1 - \lambda_k)\mathbf{z}^k + \lambda_k \mathbf{x}_f^k$;

The relaxed variable metric FBS algorithm can be applied whenever \mathbf{g} is differentiable and $\nabla \mathbf{g}$ is $(1/\beta)$ -Lipschitz for some $\beta > 0$.

Algorithm 3: Relaxed variable metric forward-backward algorithm (FBS)

input : $\mathbf{z}^0 \in \mathbf{H}; \rho > 0; \varepsilon \in (0, 2\beta\rho)$;
 $(\gamma_j)_{j \in \mathbf{N}} \subseteq (0, 2\beta\rho - \varepsilon]$;
 $\alpha_k := (2\beta\rho)/(4\beta\rho - \gamma_k)$ for $k \in \mathbf{N}$;
 $\delta \in (0, \inf\{1/\alpha_j \mid j \in \mathbf{N}\})$; *//comment: this interval is nonempty*
 $\lambda_k \in (0, 1/\alpha_k - \delta]$ for $k \in \mathbf{N}$;
 $(U_j)_{j \in \mathbf{N}} \subseteq \mathcal{S}_\rho(\mathbf{H})$.
for $k = 0, 1, \dots$ **do**
 $\mathbf{x}_g^k = \mathbf{z}^k$;
 $\mathbf{x}_f^k = J_{\gamma_k U_k^{-1}(\partial \mathbf{f} + \mathbf{S})}(\mathbf{z}^k - \gamma_k U_k^{-1} \nabla \mathbf{g}(\mathbf{z}^k))$;
 $\mathbf{z}^{k+1} = (1 - \lambda_k)\mathbf{z}^k + \lambda_k \mathbf{x}_f^k$;

In the relaxed PRS algorithm, we fix the metric and the implicit stepsize parameters throughout the course of the algorithm. We do this because the fixed-points of the PRS operator can vary with γ and U . Thus, changing these parameters will lead to an algorithm that “chases” a new fixed-point at each iteration.

Algorithm 4: Relaxed Peaceman-Rachford splitting (PRS)

input : $\mathbf{z}^0 \in \mathbf{H}; (\lambda_j)_{j \in \mathbf{N}} \subseteq (0, 2]; \gamma > 0; \rho > 0; U \in \mathcal{S}_\rho(\mathbf{H}); w \in \mathbf{R}$.
for $k = 0, 1, \dots$ **do**
 $\mathbf{z}^{k+1} = (1 - \frac{\lambda_k}{2})\mathbf{z}^k + \frac{\lambda_k}{2} \mathbf{refl}_{\gamma U^{-1}(\partial \mathbf{f} + w\mathbf{S})} \circ \mathbf{refl}_{\gamma U^{-1}(\partial \mathbf{g} + (1-w)\mathbf{S})}(\mathbf{z}^k)$;

The variable metric FBF algorithm can be applied whenever \mathbf{g} is differentiable

and $\nabla \mathbf{g}$ is $(1/\beta)$ -Lipschitz for some $\beta > 0$.

Algorithm 5: Variable metric forward-backward-forward algorithm (FBF)

input : $\mathbf{z}^0 \in \mathbf{H}; \rho > 0; (U_j)_{j \in \mathbf{N}} \subseteq \mathcal{S}_\rho(\mathbf{H}); (\gamma_j)_{j \in \mathbf{N}} \subseteq (0, \rho / (\beta^{-1} + \|\mathbf{S}\|))$.
for $k = 0, 1, \dots$ **do**
 $\mathbf{y}^k = \mathbf{z}^k - \gamma_k U_k^{-1}(\nabla \mathbf{g}(\mathbf{z}^k) + \mathbf{S} \mathbf{z}^k);$
 $\mathbf{x}_f^k = J_{\gamma_k U_k^{-1} \partial \mathbf{f}}(\mathbf{y}^k);$
 $\mathbf{w}^k = \mathbf{x}_f^k - \gamma_k U_k^{-1}(\nabla \mathbf{g}(\mathbf{x}_f^k) + \mathbf{S} \mathbf{x}_f^k);$
 $\mathbf{z}^{k+1} = \mathbf{z}^k - \mathbf{y}^k + \mathbf{w}^k;$

The following lemma relates the above algorithms to the unifying scheme.

LEMMA 2.1. *Algorithms 2, 3, 4, and 5 are special cases of the unifying scheme. In particular, the following hold for all $k \in \mathbf{N}$:*

1. In Algorithm 2, we have $\mathbf{x}_g^k := \mathbf{z}^k$, $\mathbf{x}_S^k := \mathbf{x}_f^k$, and

$$\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) := (1/\gamma_k) U_k(\mathbf{z}^k - \mathbf{x}_f^k) - \mathbf{S} \mathbf{x}_f^k \in \partial \mathbf{f}(\mathbf{x}_f^k).$$

2. In Algorithm 3, we have $\mathbf{x}_g^k := \mathbf{z}^k$, $\mathbf{x}_S^k := \mathbf{x}_f^k$, and

$$\tilde{\nabla} f(\mathbf{x}_f^k) := (1/\gamma_k) U_k(\mathbf{z}^k - \gamma_k U_k^{-1} \nabla g(\mathbf{z}^k) - \mathbf{x}_f^k) - \mathbf{S} \mathbf{x}_f^k \in \partial \mathbf{f}(\mathbf{x}_f^k).$$

3. In Algorithm 4, we have $\mathbf{z}^{k+1} - \mathbf{z}^k = \lambda_k(\mathbf{x}_f^k - \mathbf{x}_g^k)$ for

$$\begin{aligned} \mathbf{x}_g^k &:= J_{\gamma U^{-1}(\partial \mathbf{g} + (1-w)\mathbf{S})}(\mathbf{z}^k); & \mathbf{x}_f^k &:= J_{\gamma U^{-1}(\partial \mathbf{f} + w\mathbf{S})} \circ \text{refl}_{\gamma U^{-1}(\partial \mathbf{g} + (1-w)\mathbf{S})}(\mathbf{z}^k); \\ \mathbf{x}_S^k &:= w \mathbf{x}_f^k + (1-w) \mathbf{x}_g^k; & \tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) &:= (1/\gamma) U(\mathbf{z}^k - \mathbf{x}_g^k) - (1-w) \mathbf{S} \mathbf{x}_g^k \in \partial \mathbf{g}(\mathbf{x}_g^k); \end{aligned}$$

and $\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) := (1/\gamma) U(2\mathbf{x}_g^k - \mathbf{z}^k - \mathbf{x}_f^k) - w \mathbf{S} \mathbf{x}_f^k \in \partial \mathbf{f}(\mathbf{x}_f^k)$.

4. In Algorithm 5, we have $\lambda_k = 1$, $\mathbf{x}_g^k := \mathbf{x}_f^k$, $\mathbf{x}_S^k := \mathbf{x}_f^k$, and

$$\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) := (1/\gamma_k) U_k(\mathbf{y}^k - \mathbf{x}_f^k) \in \partial \mathbf{f}(\mathbf{x}_f^k).$$

Proof. Fix $k \in \mathbf{N}$, and note that the subgradient identities all follow from Part 1 of Proposition 1.2.

Part 1: This is immediate.

Part 2: From Part 1 of Proposition 1.2, we have the following identity:

$$\mathbf{x}_f^k = \mathbf{z}^k - \gamma_k U_k^{-1} \left(\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) + \nabla \mathbf{g}(\mathbf{x}_g^k) + \mathbf{S} \mathbf{x}_S^k \right).$$

Thus, altogether we have $\mathbf{z}^{k+1} = \mathbf{z}^k - \gamma_k \lambda_k U_k^{-1} \left(\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) + \nabla \mathbf{g}(\mathbf{x}_g^k) + \mathbf{S} \mathbf{x}_S^k \right)$.

Part 3: We have

$$\begin{aligned} & \text{refl}_{\gamma U^{-1}(\partial \mathbf{f} + w\mathbf{S})} \circ \text{refl}_{\gamma U^{-1}(\partial \mathbf{g} + (1-w)\mathbf{S})}(\mathbf{z}^k) \\ &= \text{refl}_{\gamma U^{-1}(\partial \mathbf{f} + w\mathbf{S})}(\mathbf{z}^k - 2\gamma U^{-1}(\tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) + (1-w)\mathbf{S} \mathbf{x}_g^k)) \\ &= \mathbf{z}^k - 2\gamma U^{-1}(\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) + \tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) + \mathbf{S}(w \mathbf{x}_f^k + (1-w) \mathbf{x}_g^k)). \end{aligned}$$

Therefore, if we define $\mathbf{x}_S^k := w \mathbf{x}_f^k + (1-w) \mathbf{x}_g^k$, then

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \gamma \lambda_k U^{-1} \left(\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) + \tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) + \mathbf{S} \mathbf{x}_S^k \right).$$

Part 4: We have

$$\mathbf{z}^{k+1} - \mathbf{z}^k = \mathbf{w}^k - \mathbf{y}^k = \mathbf{w}^k - \mathbf{x}_f^k + \mathbf{x}_f^k - \mathbf{y}^k = -\gamma_k U_k^{-1} \left(\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) + \nabla \mathbf{g}(\mathbf{x}_f^k) + \mathbf{S} \mathbf{x}_f^k \right). \square$$

2.2.1. Convergence properties. Now we establish two basic and well known results on the boundedness and summability of various terms related to the above algorithms. These facts will be used repeatedly in our convergence rate analysis.

PROPOSITION 2.2 (Averagedness properties). *Let $\rho \in \mathbf{R}_{++}$, let $U \in \mathcal{S}_\rho(\mathbf{H})$, let $\gamma \in \mathbf{R}_{++}$, and let $\beta \in \mathbf{R}_{++}$. Then the following hold:*

1. *The operator $J_{\gamma U^{-1}(\partial \mathbf{f} + \mathbf{S})}$ is $(1/2)$ -averaged in the norm $\|\cdot\|_U$. In addition, the set of fixed points of $J_{\gamma U^{-1}(\partial \mathbf{f} + \mathbf{S})}$ is equal to $\text{zer}(\partial \mathbf{f} + \mathbf{S})$*
2. *Let $\gamma \in (0, 2\beta\rho)$. Suppose that \mathbf{g} is differentiable and $\nabla \mathbf{g}$ is $(1/\beta)$ -Lipschitz. Then the composition*

$$T_{\text{FBS}}^{U,\gamma} := J_{\gamma U^{-1}(\partial \mathbf{f} + \mathbf{S})} \circ (I_{\mathbf{H}} - \gamma U^{-1} \nabla \mathbf{g}) \quad (2.2)$$

is $\alpha_{\rho,\gamma}$ -averaged in the norm $\|\cdot\|_U$ where

$$\alpha_{\rho,\gamma} := \frac{2\beta\rho}{4\beta\rho - \gamma}. \quad (2.3)$$

In addition, the set of fixed points of $T_{\text{FBS}}^{U,\gamma}$ is equal to $\text{zer}(\partial \mathbf{f} + \nabla \mathbf{g} + \mathbf{S})$.

3. *Let $w \in \mathbf{R}$, and define the PRS operator:*

$$T_{\text{PRS}} := \text{refl}_{\gamma U^{-1}(\partial \mathbf{f} + w\mathbf{S})} \circ \text{refl}_{\gamma U^{-1}(\partial \mathbf{g} + (1-w)\mathbf{S})}. \quad (2.4)$$

Then T_{PRS} is nonexpansive in the metric $\|\cdot\|_U$. Thus, the following DRS operator

$$(T_{\text{PRS}})_{1/2} = \frac{1}{2} I_{\mathbf{H}} + \frac{1}{2} \text{refl}_{\gamma U^{-1}(\partial \mathbf{f} + w\mathbf{S})} \circ \text{refl}_{\gamma U^{-1}(\partial \mathbf{g} + (1-w)\mathbf{S})} \quad (2.5)$$

is $(1/2)$ -averaged. In addition, the set of fixed points of T_{PRS} and $(T_{\text{PRS}})_{1/2}$ coincide and $\text{zer}(\partial \mathbf{f} + \partial \mathbf{g} + \mathbf{S}) = \{J_{\gamma U^{-1}(\partial \mathbf{g} + (1-w)\mathbf{S})}(\mathbf{z}) \mid \mathbf{z} \in \mathbf{H} \text{ and } T_{\text{PRS}}\mathbf{z} = \mathbf{z}\}$.

Proof. Parts 1 and 3 are simple modifications of standard facts found in [2].

Part 2: Note that $U^{-1} \nabla \mathbf{g}$ is $\beta\rho$ -cocoercive in $\|\cdot\|_U$ by Proposition 1.5 and the Baillon-Haddad theorem [1]. Thus, $I_{\mathbf{H}} - \gamma U^{-1} \nabla \mathbf{g}$ is $\gamma/(2\beta\rho)$ averaged in $\|\cdot\|_U$ by [2, Proposition 4.33]. Thus, the formula for $\alpha_{\rho,\gamma}$ follows from [36, Theorem 3(b)]. The fixed-point identity follows from a simple modification of [2, Theorem 25.1]. \square

PROPOSITION 2.3 (Bounded and summable sequences). *The following hold:*

1. *Let $\mathbf{z}^* \in \text{zer}(\partial \mathbf{f} + \mathbf{S})$. Then in Algorithm 2, we have for all $k \in \mathbf{N}$, that $\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_{U_{k+1}}^2 \leq (1 + \eta_k) \|\mathbf{z}^k - \mathbf{z}^*\|_{U_k}^2$ and hence, $\|\mathbf{z}^k - \mathbf{z}^*\|_{U_k}^2 \leq \eta_{\text{p}} \|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2$.*
2. *Let $\mathbf{z}^* \in \text{zer}(\partial \mathbf{f} + \nabla \mathbf{g} + \mathbf{S})$. Then in Algorithm 3, the following are true:*
 - (i) *For all $k \in \mathbf{N}$, $\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_{U_{k+1}}^2 \leq (1 + \eta_k) \|\mathbf{z}^k - \mathbf{z}^*\|_{U_k}^2$ and hence, $\|\mathbf{z}^k - \mathbf{z}^*\|_{U_k}^2 \leq \eta_{\text{p}} \|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2$.*
 - (ii) *The following sum is finite:*

$$\sum_{i=0}^{\infty} \frac{1 - \alpha_i \lambda_i}{\alpha_i \lambda_i} \|\mathbf{z}^{i+1} - \mathbf{z}^i\|^2 \leq \frac{1}{\rho} (1 + \eta_{\text{p}} \eta_{\text{s}}) \|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2. \quad (2.6)$$

3. *Let \mathbf{z}^* be a fixed-point of T_{PRS} . Then in Algorithm 4, we have for all $k \in \mathbf{N}$, that $\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_U^2 \leq \|\mathbf{z}^k - \mathbf{z}^*\|_U^2$ and hence, $\|\mathbf{z}^k - \mathbf{z}^*\|_U^2 \leq \|\mathbf{z}^0 - \mathbf{z}^*\|_U^2$.*

4. *Let $\mathbf{z}^* \in \text{zer}(\partial \mathbf{f} + \nabla \mathbf{g} + \mathbf{S})$. Then in Algorithm 5, we have for all $k \in \mathbf{N}$ that $\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_{U_{k+1}}^2 \leq (1 + \eta_k) \|\mathbf{z}^k - \mathbf{z}^*\|_{U_k}^2$ and hence, $\|\mathbf{z}^k - \mathbf{z}^*\|_{U_k}^2 \leq \eta_{\text{p}} \|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2$.*

Proof. Parts 1, 2, and 3 follow from Proposition 1.4 applied to the sequences of operators $(T_j)_{j \in \mathbf{N}} := (J_{\gamma_j U_j^{-1}(\partial \mathbf{f} + \mathbf{S})})_{j \in \mathbf{N}}$, $(T_j)_{j \in \mathbf{N}} := (T_{\text{FBS}}^{U_j, \gamma_j})_{j \in \mathbf{N}}$, and $(T_j)_{j \in \mathbf{N}} := ((T_{\text{PRS}})_{1/2})_{j \in \mathbf{N}}$, respectively.

Part 4 follows from Proposition 1.6 applied to the maximal monotone operator $\partial \mathbf{f}$ and the $(\beta^{-1} + \|\mathbf{S}\|)$ -Lipschitz operator $\nabla \mathbf{g} + \mathbf{S}$. \square

2.3. The fundamental inequality. This section describes the pre-primal-dual gap (Definition 2.5). We use the pre-primal-dual gap to measure the convergence of the unifying scheme. In Section 5, we will show that under certain conditions, the pre-primal-dual gap function bounds the primal and dual objective errors of the iterates generated by a class of primal-dual algorithms.

Before we introduce the gap function, we analyze the optimality conditions of Problem 1. The following lemma is well-known.

LEMMA 2.4. *Let $\mathbf{x}^* \in \mathbf{H}$. Suppose that \mathbf{x}^* solves Problem 1. Then for all $\mathbf{x} \in \mathbf{H}$,*

$$\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x}) + \langle \mathbf{S}\mathbf{x}, -\mathbf{x}^* \rangle - \mathbf{f}(\mathbf{x}^*) - \mathbf{g}(\mathbf{x}^*) \geq 0. \quad (2.7)$$

On the other hand, if $\partial(\mathbf{f} + \mathbf{g})(\mathbf{x}^) = \partial \mathbf{f}(\mathbf{x}^*) + \partial \mathbf{g}(\mathbf{x}^*)$ and \mathbf{x}^* satisfies Equation (2.7) for all $\mathbf{x} \in \text{dom}(\mathbf{f}) \cap \text{dom}(\mathbf{g})$, then \mathbf{x}^* solves Problem 1.*

Proof. If \mathbf{x}^* solves Problem 1, then $-\mathbf{S}\mathbf{x}^*$ is a subgradient of $\mathbf{f} + \mathbf{g}$ at the point \mathbf{x}^* . Thus, Equation (2.7) follows after noting that $\langle \mathbf{S}\mathbf{x}, \mathbf{x} \rangle = 0$ for all $\mathbf{x} \in \mathbf{H}$.

The other direction follows because Equation (2.7) characterizes the set of subgradients of the form $-\mathbf{S}\mathbf{x}^* \in \partial(\mathbf{f} + \mathbf{g})(\mathbf{x}^*) = \partial \mathbf{f}(\mathbf{x}^*) + \partial \mathbf{g}(\mathbf{x}^*)$. \square

See [2, Corollary 16.38] for conditions that imply additivity of the subdifferential.

Lemma 2.4 motivates the following definition:

DEFINITION 2.5 (Pre-primal-dual gap). *Let the setting be as in Algorithm 1. Define the pre-primal dual gap function by the formula: for all $\mathbf{x}_f, \mathbf{x}_g, \mathbf{x}_S, \mathbf{x} \in \mathbf{H}$, let*

$$\mathcal{G}^{\text{pre}}(\mathbf{x}_f, \mathbf{x}_g, \mathbf{x}_S; \mathbf{x}) = \mathbf{f}(\mathbf{x}_f) + \mathbf{g}(\mathbf{x}_g) + \langle \mathbf{S}\mathbf{x}_S, -\mathbf{x} \rangle - \mathbf{f}(\mathbf{x}) - \mathbf{g}(\mathbf{x}). \quad (2.8)$$

We name \mathcal{G}^{pre} the pre-primal-dual-gap function after the standard primal-dual gap function that appears in [15, 6, 11]. We use the word “pre” because the standard primal-dual gap function usually involves a supremum over the last variable \mathbf{x} . Note that if $\partial(\mathbf{f} + \mathbf{g})(\mathbf{x}') = \partial \mathbf{f}(\mathbf{x}') + \partial \mathbf{g}(\mathbf{x}')$ and

$$\sup_{\mathbf{x} \in \mathbf{H}} \mathcal{G}^{\text{pre}}(\mathbf{x}', \mathbf{x}', \mathbf{x}'; \mathbf{x}) \leq 0, \quad (2.9)$$

then \mathbf{x}' is a solution of Problem 1 (Lemma 2.4).

Our goal throughout the rest of this paper is to bound the pre-primal-dual gap when $\mathbf{x}_f = \mathbf{x}_g = \mathbf{x}_S$. Because of Equation (2.9), all of our upper bounds will be a function of the norm of the last component of \mathcal{G}^{pre} . In some cases, we can restrict the supremum in Equation (2.9) to a smaller subset $C \subseteq \mathbf{H}$. This is the case if, for example, $\text{dom}(\mathbf{f}) \cap \text{dom}(\mathbf{g})$ is bounded. Whenever the supremum can be restricted, we obtain a meaningful convergence rate.

Finally, Lemma 2.4 shows that for all $\mathbf{x} \in \mathbf{H}$,

$$\mathcal{G}^{\text{pre}}(\mathbf{x}, \mathbf{x}, \mathbf{x}; \mathbf{x}^*) \geq 0 \quad (2.10)$$

whenever \mathbf{x}^* solves Problem 1. See Section 5.1 for other lower bounds of the pre-primal-dual gap in the context of a particular convex optimization problem.

The following is our main tool to bound the pre-primal-dual gap.

PROPOSITION 2.6 (Upper fundamental inequality for primal dual schemes). *Suppose that $(\mathbf{z}^j)_{j \geq 0}$ is generated by Algorithm 1, and let $\mathbf{x} \in \mathbf{H}$. Then the following inequality holds: for all $k \in \mathbf{N}$,*

$$\begin{aligned} 2\gamma_k \lambda_k \mathcal{G}^{\text{pre}}(\mathbf{x}_f^k, \mathbf{x}_g^k, \mathbf{x}_s^k; \mathbf{x}) &\leq \|\mathbf{z}^k - \mathbf{x}\|_{U_k}^2 - \|\mathbf{z}^{k+1} - \mathbf{x}\|_{U_k}^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^k\|_{U_k}^2 \\ &\quad + 2\gamma_k \lambda_k \langle \mathbf{x}_f^k - \mathbf{z}^{k+1}, \tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) \rangle \\ &\quad + 2\gamma_k \lambda_k \langle \mathbf{x}_g^k - \mathbf{z}^{k+1}, \tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) \rangle \\ &\quad + 2\gamma_k \lambda_k \langle -\mathbf{z}^{k+1}, \mathbf{S} \mathbf{x}_s^k \rangle. \end{aligned} \quad (2.11)$$

Proof. Fix $k \in \mathbf{N}$. First expand the norm:

$$\|\mathbf{z}^{k+1} - \mathbf{x}\|_{U_k}^2 = \|\mathbf{z}^k - \mathbf{x}\|_{U_k}^2 + 2\langle \mathbf{x} - \mathbf{z}^{k+1}, \mathbf{z}^k - \mathbf{z}^{k+1} \rangle_{U_k} - \|\mathbf{z}^{k+1} - \mathbf{z}^k\|_{U_k}^2.$$

Now we expand the inner product:

$$\begin{aligned} 2\langle \mathbf{x} - \mathbf{z}^{k+1}, \mathbf{z}^k - \mathbf{z}^{k+1} \rangle_{U_k} &= 2\langle \mathbf{x} - \mathbf{z}^{k+1}, \gamma_k \lambda_k U_k^{-1} (\tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) + \tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) + \mathbf{S} \mathbf{x}_s^k) \rangle_{U_k} \\ &= 2\gamma_k \lambda_k \langle \mathbf{x} - \mathbf{z}^{k+1}, \tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) \rangle + 2\gamma_k \lambda_k \langle \mathbf{x} - \mathbf{z}^{k+1}, \tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) \rangle \\ &\quad + 2\gamma_k \lambda_k \langle \mathbf{x} - \mathbf{z}^{k+1}, \mathbf{S} \mathbf{x}_s^k \rangle. \end{aligned}$$

We add and subtract a point in the inner products involving \mathbf{f} and \mathbf{g} and use the subgradient inequality to get:

$$\begin{aligned} 2\gamma_k \lambda_k \langle \mathbf{x} - \mathbf{z}^{k+1}, \tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) \rangle &\leq 2\gamma_k \lambda_k \langle \mathbf{x}_f^k - \mathbf{z}^{k+1}, \tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) \rangle + 2\gamma_k \lambda_k (\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_f^k)); \\ 2\gamma_k \lambda_k \langle \mathbf{x} - \mathbf{z}^{k+1}, \tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) \rangle &\leq 2\gamma_k \lambda_k \langle \mathbf{x}_g^k - \mathbf{z}^{k+1}, \tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) \rangle + 2\gamma_k \lambda_k (\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}_g^k)). \end{aligned}$$

Therefore Equation (2.11) follows after rearranging. \square

The upper fundamental inequality in Proposition 2.6 bounds the pre-primal-dual gap with the sum of an alternating sequence and a key term.

DEFINITION 2.7 (Upper key term). *Let $(\mathbf{z}^j)_{j \in \mathbf{N}}$ be generated by Algorithm 1. For all $k \in \mathbf{N}$, we define the fundamental upper key term*

$$\begin{aligned} \kappa_u^k(\lambda_k) &:= -\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_{U_k}^2 \\ &\quad + 2\gamma_k \lambda_k \langle \mathbf{x}_f^k - \mathbf{z}^{k+1}, \tilde{\nabla} \mathbf{f}(\mathbf{x}_f^k) \rangle \\ &\quad + 2\gamma_k \lambda_k \langle \mathbf{x}_g^k - \mathbf{z}^{k+1}, \tilde{\nabla} \mathbf{g}(\mathbf{x}_g^k) \rangle \\ &\quad + 2\gamma_k \lambda_k \langle -\mathbf{z}^{k+1}, \mathbf{S} \mathbf{x}_s^k \rangle. \end{aligned} \quad (2.12)$$

The value $\kappa_u^k(\lambda_k)$ depends on the entire history of Algorithm 1 up to and including iteration k , but in our analysis we will only view $\kappa_u^k(\lambda_k)$ as a function of the parameter λ_k . Throughout the rest of the paper, we will often make the dependence of the upper key term on λ_k implicit, and denote $\kappa_u^k := \kappa_u^k(\lambda_k)$. However, in the proof of Theorem 4 we will need to keep the dependence explicit.

2.3.1. Computing the upper key terms. The following proposition will compute the upper key terms induced by the PPA, FBS, PRS, and FBF algorithms. See Section 2.2 for the definitions of the points \mathbf{x}_f^k , \mathbf{x}_g^k , and \mathbf{x}_s^k .

PROPOSITION 2.8 (Computing the upper key terms). *Let $(\mathbf{z}^j)_{j \in \mathbf{N}}$ be generated by Algorithm 1. Then for all $k \in \mathbf{N}$, the following inequalities and identities hold:*

1. In Algorithm 2, we have $\kappa_u^k(\lambda_k) = (1 - 2/\lambda_k) \|\mathbf{z}^{k+1} - \mathbf{z}^k\|_{U_k}^2$.
2. In Algorithm 3, we have

$$\kappa_u^k(\lambda_k) \leq \left(\rho - \frac{\varepsilon}{\beta \lambda_k} \right) \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + 2\gamma_k \lambda_k \mathbf{g}(\mathbf{x}_g^k) - 2\gamma_k \lambda_k \mathbf{g}(\mathbf{x}_f^k).$$

3. In Algorithm 4, we have $\kappa_u^k(\lambda_k) = (1 - 2/\lambda_k) \|\mathbf{z}^{k+1} - \mathbf{z}^k\|_{U_k}^2$.
4. In Algorithm 5, we have $\kappa_u^k(\lambda_k) \leq 0$.

Proof. Fix $k \in \mathbf{N}$. To simplify notation, we drop the iteration index and denote $\mathbf{z} := \mathbf{z}^k$, $\mathbf{x}_f := \mathbf{x}_f^k$, $\mathbf{x}_g := \mathbf{x}_g^k$, $\mathbf{x}_S := \mathbf{x}_S^k$, $\mathbf{z}^+ := \mathbf{z}^{k+1}$, $\gamma := \gamma_k$, $\lambda := \lambda_k$, $U := U_k$, and $\kappa_u := \kappa_u^k(\lambda_k)$ throughout this proof.

For PPA, FBS, and PRS, we note that the following identities hold:

$$\mathbf{z}^+ - \mathbf{z} = \lambda(\mathbf{x}_f - \mathbf{x}_g), \quad (2.13)$$

and there exists $w \in \mathbf{R}$ such that

$$\mathbf{x}_S = w\mathbf{x}_f + (1 - w)\mathbf{x}_g. \quad (2.14)$$

Indeed, in PPA and FBS, $w = 1$ (see Section 2.2). In PRS, w is a parameter of the algorithm, and Equations (2.14) and (2.13) are shown in Lemma 2.1. Furthermore, Part 1 of Proposition 1.2 shows that in PPA and FBS,

$$\mathbf{x}_f = \mathbf{x}_g - \gamma U^{-1} \left(\tilde{\nabla} \mathbf{f}(\mathbf{x}_f) + \nabla \mathbf{g}(\mathbf{x}_g) + \mathbf{S} \mathbf{x}_S \right) \quad (2.15)$$

for a unique subgradient $\tilde{\nabla} \mathbf{f}(\mathbf{x}_f) \in \partial \mathbf{f}(\mathbf{x}_f)$; see Lemma 2.1 for the definition of $\tilde{\nabla} \mathbf{f}(\mathbf{x}_f)$.

Now we claim that in PPA, FBS, and PRS,

$$\kappa_u = 2\langle \mathbf{x}_f + \gamma U^{-1}(\tilde{\nabla} \mathbf{g}(\mathbf{x}_g) + (1 - w)\mathbf{S} \mathbf{x}_g) - \mathbf{z}^+, \mathbf{z} - \mathbf{z}^+ \rangle_U - \|\mathbf{z}^+ - \mathbf{z}\|_U^2 \quad (2.16)$$

where we make the identification $\tilde{\nabla} \mathbf{g}(\mathbf{x}_g) = \nabla \mathbf{g}(\mathbf{x}_g)$ whenever \mathbf{g} is differentiable; see Lemma 2.1 for the definition of $\tilde{\nabla} \mathbf{g}(\mathbf{x}_g)$ in the PRS algorithm. Because $\mathbf{x}_S = \mathbf{x}_g + w(\mathbf{x}_f - \mathbf{x}_g) = \mathbf{x}_g + (w/\lambda)(\mathbf{z}^+ - \mathbf{z})$ and $\langle \mathbf{S} \mathbf{x}, \mathbf{x} \rangle = 0$ for all $\mathbf{x} \in \mathbf{H}$, we have the simplification:

$$2\langle \mathbf{z} - \mathbf{z}^+, \gamma(1 - w)\mathbf{S} \mathbf{x}_S \rangle = 2\langle \mathbf{z} - \mathbf{z}^+, \gamma(1 - w)\mathbf{S} \mathbf{x}_g \rangle. \quad (2.17)$$

Therefore,

$$\begin{aligned} \kappa_u &= -\|\mathbf{z}^+ - \mathbf{z}\|_U^2 + 2\gamma\lambda\langle \mathbf{x}_f - \mathbf{z}^+, \tilde{\nabla} \mathbf{f}(\mathbf{x}_f) \rangle \\ &\quad + 2\gamma\lambda\langle \mathbf{x}_g - \mathbf{z}^+, \tilde{\nabla} \mathbf{g}(\mathbf{x}_g) \rangle + 2\gamma\lambda\langle \mathbf{x}_S - \mathbf{z}^+, \mathbf{S} \mathbf{x}_S \rangle \\ &= -\|\mathbf{z}^+ - \mathbf{z}\|_U^2 + 2\gamma\lambda\langle \mathbf{x}_f - \mathbf{z}^+, \tilde{\nabla} \mathbf{f}(\mathbf{x}_f) \rangle \\ &\quad + 2\gamma\lambda\langle \mathbf{x}_g - \mathbf{x}_f, \tilde{\nabla} \mathbf{g}(\mathbf{x}_g) \rangle + 2\gamma\lambda\langle \mathbf{x}_f - \mathbf{z}^+, \tilde{\nabla} \mathbf{g}(\mathbf{x}_g) \rangle \\ &\quad + 2\gamma\lambda\langle \mathbf{x}_S - \mathbf{x}_f, \mathbf{S} \mathbf{x}_S \rangle + 2\gamma\lambda\langle \mathbf{x}_f - \mathbf{z}^+, \mathbf{S} \mathbf{x}_S \rangle \\ &= -\|\mathbf{z}^+ - \mathbf{z}\|_U^2 + 2\gamma\lambda\langle \mathbf{x}_f - \mathbf{z}^+, \tilde{\nabla} \mathbf{f}(\mathbf{x}_f) + \tilde{\nabla} \mathbf{g}(\mathbf{x}_g) + \mathbf{S} \mathbf{x}_S \rangle \\ &\stackrel{(2.13)}{=} -\|\mathbf{z}^+ - \mathbf{z}\|_U^2 + 2\langle \mathbf{z} - \mathbf{z}^+, \gamma \tilde{\nabla} \mathbf{g}(\mathbf{x}_g) \rangle + 2\langle \mathbf{z} - \mathbf{z}^+, \gamma(1 - w)\mathbf{S} \mathbf{x}_S \rangle \\ &\stackrel{(2.13)}{=} -\|\mathbf{z}^+ - \mathbf{z}\|_U^2 + 2\langle \mathbf{x}_f - \mathbf{z}^+, \mathbf{z} - \mathbf{z}^+ \rangle_U \\ &\stackrel{(2.17)}{=} -\|\mathbf{z}^+ - \mathbf{z}\|_U^2 + 2\langle \mathbf{z} - \mathbf{z}^+, \gamma \tilde{\nabla} \mathbf{g}(\mathbf{x}_g) + \gamma(1 - w)\mathbf{S} \mathbf{x}_g \rangle \\ &= 2\langle \mathbf{x}_f + \gamma U^{-1}(\tilde{\nabla} \mathbf{g}(\mathbf{x}_g) + (1 - w)\mathbf{S} \mathbf{x}_g) - \mathbf{z}^+, \mathbf{z} - \mathbf{z}^+ \rangle_U - \|\mathbf{z}^+ - \mathbf{z}\|_U^2 \end{aligned}$$

where the second to last equality uses Equation (2.15) and the second to last “+” also uses Equation (2.14).

Now we proceed with the specific cases: In PPA and FBS, $w = 1$ and

$$\begin{aligned}
\kappa_u &\stackrel{(2.16)}{=} 2\langle \mathbf{x}_f + \gamma U^{-1} \nabla \mathbf{g}(\mathbf{x}_g) - \mathbf{z}^+, \mathbf{z} - \mathbf{z}^+ \rangle_U - \|\mathbf{z}^+ - \mathbf{z}\|_U^2 \\
&= 2\langle \mathbf{x}_f - \mathbf{z}^+, \mathbf{z} - \mathbf{z}^+ \rangle_U + 2\gamma \langle \nabla \mathbf{g}(\mathbf{x}_g), \mathbf{z} - \mathbf{z}^+ \rangle - \|\mathbf{z}^+ - \mathbf{z}\|_U^2 \\
&= 2 \left(1 - \frac{1}{\lambda} \right) \|\mathbf{z}^+ - \mathbf{z}\|_U^2 + 2\gamma \lambda \langle \nabla \mathbf{g}(\mathbf{x}_g), \mathbf{x}_g - \mathbf{x}_f \rangle - \|\mathbf{z}^+ - \mathbf{z}\|_U^2 \\
&\leq \left(1 - \frac{2}{\lambda} \right) \|\mathbf{z}^+ - \mathbf{z}\|_U^2 + 2\gamma \lambda \mathbf{g}(\mathbf{x}_g) - 2\gamma \lambda \mathbf{g}(\mathbf{x}_f) + \frac{\gamma}{\lambda \beta} \|\mathbf{z}^+ - \mathbf{z}\|^2
\end{aligned} \tag{2.18}$$

where we use the identity $\mathbf{x}_f - \mathbf{z}^+ = (1 - (1/\lambda))(\mathbf{z} - \mathbf{z}^+)$ on the third line, we use the identity $\mathbf{z}^+ - \mathbf{z} = \lambda(\mathbf{x}_f - \mathbf{x}_g)$ (Equation (2.13)) on the last two lines, and the last inequality follows from the Descent Theorem [2, Theorem 18.15(iii)]: $\langle \nabla \mathbf{g}(\mathbf{x}_g), \mathbf{x}_g - \mathbf{x}_f \rangle \leq \mathbf{g}(\mathbf{x}_g) - \mathbf{g}(\mathbf{x}_f) + (1/(2\beta))\|\mathbf{x}_g - \mathbf{x}_f\|^2$. In PPA $\mathbf{g} \equiv 0$, so the Equation (2.18) implies the identity in Part 1. The inequality for FBS now follows by the above bound for κ_u , the bound $\gamma \leq 2\beta\rho - \varepsilon$, and

$$\left(1 - \frac{2}{\lambda} \right) \|\mathbf{z}^+ - \mathbf{z}\|_U^2 + \frac{\gamma}{\lambda \beta} \|\mathbf{z}^+ - \mathbf{z}\|^2 \leq \left(\rho + \frac{\gamma - 2\beta\rho}{\lambda \beta} \right) \|\mathbf{z}^+ - \mathbf{z}\|^2$$

where we use $\lambda \leq (4\beta\rho - \gamma)/2\beta\rho \leq 2$ and the lower bound $U \succcurlyeq \rho I_{\mathbf{H}}$.

For relaxed PRS, we have

$$\begin{aligned}
\mathbf{z}^+ &= \mathbf{z} + \lambda(\mathbf{x}_f - \mathbf{x}_g) = (1 - \lambda)\mathbf{z} + \lambda(\mathbf{x}_f - \mathbf{x}_g + \mathbf{z}) \\
&= (1 - \lambda)\mathbf{z} + \lambda \left(\mathbf{x}_f + \gamma U^{-1} \left(\tilde{\nabla} \mathbf{g}(\mathbf{x}_g) + (1 - w)\mathbf{S}\mathbf{x}_g \right) \right).
\end{aligned}$$

Therefore, subtract $\lambda\mathbf{z}^+ + (1 - \lambda)\mathbf{z}$ from both sides of the above equation, divide by λ , and use the identity in Equation (2.16) to get

$$\begin{aligned}
\kappa_u &= 2\langle \mathbf{x}_f + \gamma U^{-1} \left(\tilde{\nabla} \mathbf{g}(\mathbf{x}_g) + (1 - w)\mathbf{S}\mathbf{x}_g \right) - \mathbf{z}^+, \mathbf{z} - \mathbf{z}^+ \rangle_U - \|\mathbf{z}^+ - \mathbf{z}\|_U^2 \\
&= 2 \left(1 - \frac{1}{\lambda} \right) \|\mathbf{z}^+ - \mathbf{z}\|_U^2 - \|\mathbf{z}^+ - \mathbf{z}\|_U^2 = \left(1 - \frac{2}{\lambda} \right) \|\mathbf{z}^+ - \mathbf{z}\|_U^2.
\end{aligned}$$

Finally, we prove the bound for the FBF algorithm:

$$\begin{aligned}
\kappa_u &\stackrel{(2.12)}{=} 2\langle \mathbf{x}_f - \mathbf{z}^+, \gamma \tilde{\nabla} \mathbf{f}(\mathbf{x}_f) + \gamma \nabla \mathbf{g}(\mathbf{x}_f) + \gamma \mathbf{S}\mathbf{x}_f \rangle - \|\mathbf{z}^+ - \mathbf{z}\|_U^2 \\
&= 2\langle \mathbf{x}_f - \mathbf{z}^+, \mathbf{z} - \mathbf{z}^+ \rangle_U - \|\mathbf{z}^+ - \mathbf{z}\|_U^2 \stackrel{(1.7)}{=} \|\mathbf{x}_f - \mathbf{z}^+\|_U^2 - \|\mathbf{x}_f - \mathbf{z}\|_U^2.
\end{aligned}$$

Furthermore, the identity holds:

$$\begin{aligned}
\mathbf{z}^+ - \mathbf{x}_f &= \mathbf{z}^+ - \mathbf{z} + \mathbf{z} - \mathbf{x}_f \\
&= -\gamma U^{-1} \left(\tilde{\nabla} \mathbf{f}(\mathbf{x}_f) + \nabla \mathbf{g}(\mathbf{x}_f) + \mathbf{S}\mathbf{x}_f \right) + \gamma U^{-1} \left(\tilde{\nabla} \mathbf{f}(\mathbf{x}_f) + \nabla \mathbf{g}(\mathbf{z}) + \mathbf{S}\mathbf{z} \right) \\
&= \gamma U^{-1} (\nabla \mathbf{g}(\mathbf{z}) + \mathbf{S}\mathbf{z} - \nabla \mathbf{g}(\mathbf{x}_f) - \mathbf{S}\mathbf{x}_f).
\end{aligned} \tag{2.19}$$

Note that the operator $\nabla \mathbf{g} + \mathbf{S}$ is $(1/\beta) + \|\mathbf{S}\|$ Lipschitz. Thus,

$$\begin{aligned}
\|\mathbf{x}_f - \mathbf{z}^+\|_U^2 - \|\mathbf{x}_f - \mathbf{z}\|_U^2 &\stackrel{(2.19)}{=} \gamma^2 \|\nabla \mathbf{g}(\mathbf{z}) + \mathbf{S}\mathbf{z} - \nabla \mathbf{g}(\mathbf{x}_f) - \mathbf{S}\mathbf{x}_f\|_{U^{-1}}^2 - \|\mathbf{x}_f - \mathbf{z}\|_U^2 \\
&\leq \left(\frac{\gamma^2}{\rho} \left(\frac{1}{\beta} + \|\mathbf{S}\| \right)^2 - \rho \right) \|\mathbf{x}_f - \mathbf{z}\|^2 \leq 0,
\end{aligned}$$

where we use the following bound: for all $\mathbf{x} \in \mathbf{H}$, $\|\mathbf{x}\|_{U^{-1}}^2 \leq (1/\rho)\|\mathbf{x}\|^2$ (Lemma 1.1). \square

3. Ergodic convergence. In this section, we prove an ergodic convergence rate for the pre-primal-dual gap. To this end, we recall the partial sum sequence $\Sigma_k = \sum_{i=0}^k \gamma_i \lambda_i$, and for every sequence of vectors $(\mathbf{x}^j)_{j \geq 0} \subseteq \mathbf{H}$, we define the ergodic sequence $\bar{\mathbf{x}}^k = (1/\Sigma_k) \sum_{i=0}^k \gamma_i \lambda_i \mathbf{x}^i$. For each algorithm, Theorem 3.2 (below) gives an ergodic sequence $(\bar{\mathbf{x}}^j)_{j \in \mathbf{N}}$ such that for all bounded subsets $D \subseteq \mathbf{H}$, we have

$$\sup_{\mathbf{x} \in D} \mathcal{G}^{\text{pre}}(\bar{\mathbf{x}}^k, \bar{\mathbf{x}}^k, \bar{\mathbf{x}}^k; \mathbf{x}) = O\left(\frac{1 + \sup_{\mathbf{x} \in D} \|\mathbf{x}\|^2}{\Sigma_k}\right).$$

This bound is a generalization of the primal-dual gap bounds shown in [15, 7, 6, 11]. See Section 5.1 for several lower bounds of the pre-primal-dual gap.

Before we prove our ergodic rates, we need to prove a bound for PRS. Recall that we only analyze the PRS algorithm when the map $U_k \equiv U$ is fixed. The following lemma will help us deduce the convergence rate of the PRS algorithm whenever \mathbf{f} or \mathbf{g} is Lipschitz (Part 3 of Theorem 3.2).

LEMMA 3.1. *Suppose that $(\mathbf{z}^j)_{j \in \mathbf{N}}$ is generated by the relaxed PRS algorithm and that \mathbf{z}^* is a fixed-point of T_{PRS} (see equation (2.4)). Then the following ergodic bound holds: for all $k \in \mathbf{N}$, we have*

$$\|\bar{\mathbf{x}}_{\mathbf{f}}^k - \bar{\mathbf{x}}_{\mathbf{g}}^k\|_U \leq \frac{2\gamma\|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\Sigma_k}. \quad (3.1)$$

Proof. Fix $k \in \mathbf{N}$. The identity $\lambda_k(\mathbf{x}_{\mathbf{f}}^k - \mathbf{x}_{\mathbf{g}}^k) = \mathbf{z}^{k+1} - \mathbf{z}^k$ and the fact the sequence $(\|\mathbf{z}^j - \mathbf{z}^*\|_U)_{j \in \mathbf{N}}$ is decreasing (Part 3 of Proposition 2.3), show that

$$\begin{aligned} \|\bar{\mathbf{x}}_{\mathbf{f}}^k - \bar{\mathbf{x}}_{\mathbf{g}}^k\|_U &= \left\| \frac{\gamma}{\Sigma_k} \sum_{i=0}^k \lambda_i (\mathbf{x}_{\mathbf{f}}^i - \mathbf{x}_{\mathbf{g}}^i) \right\|_U = \frac{\gamma\|\mathbf{z}^{k+1} - \mathbf{z}^0\|_U}{\Sigma_k} \leq \frac{\gamma\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_U + \gamma\|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\Sigma_k} \\ &\leq \frac{2\gamma\|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\Sigma_k}. \quad \square \end{aligned}$$

Lemma 3.1 shows that the difference of splitting variables $\bar{\mathbf{x}}_{\mathbf{f}}^k - \bar{\mathbf{x}}_{\mathbf{g}}^k$ converges to zero with rate $O(1/\Sigma_k)$. Thus, if \mathbf{f} is Lipschitz continuous, then $|\mathbf{f}(\bar{\mathbf{x}}_{\mathbf{f}}^k) - \mathbf{f}(\bar{\mathbf{x}}_{\mathbf{g}}^k)| = O(1/\Sigma_k)$.

We are now ready to prove our main ergodic convergence results.

THEOREM 3.2 (Ergodic convergence of the unifying scheme). *Suppose that the sequence $(\mathbf{z}^j)_{j \in \mathbf{N}}$ is generated by Algorithm 1, and suppose that Assumption 3 holds. Then for all $\mathbf{x} \in \mathbf{H}$ and all $k \in \mathbf{N}$, we have the following bounds:*

1. **Ergodic convergence of PPA:** Let $\mathbf{z}^* \in \text{zer}(\partial\mathbf{f} + \mathbf{S})$. Then in Algorithm 2, we have

$$\mathcal{G}^{\text{pre}}(\bar{\mathbf{x}}_{\mathbf{f}}^k, \bar{\mathbf{x}}_{\mathbf{f}}^k, \bar{\mathbf{x}}_{\mathbf{f}}^k; \mathbf{x}) \leq \frac{\|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + 2\eta_{\text{p}}\eta_{\text{s}}\|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2 + 2\mu\eta_{\text{s}}\|\mathbf{z}^* - \mathbf{x}\|^2}{2\Sigma_k}.$$

2. **Ergodic convergence of FBS:** Let $\mathbf{z}^* \in \text{zer}(\partial\mathbf{f} + \nabla\mathbf{g} + \mathbf{S})$, and let $\bar{\lambda} = \sup_{j \in \mathbf{N}} \lambda_j$. Then in Algorithm 3, we have the bounds $0 < \inf_{j \in \mathbf{N}} \lambda_j \leq \bar{\lambda} \leq 2$ and $\inf_{j \in \mathbf{N}} (1 - \alpha_j \lambda_j) / (\alpha_j \lambda_j) > 0$, and

$$\begin{aligned} &\mathcal{G}^{\text{pre}}(\bar{\mathbf{x}}_{\mathbf{f}}^k, \bar{\mathbf{x}}_{\mathbf{f}}^k, \bar{\mathbf{x}}_{\mathbf{f}}^k; \mathbf{x}) \\ &\leq \frac{\left(\|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + \left(2\eta_{\text{p}}\eta_{\text{s}} + \frac{(1+\eta_{\text{p}}\eta_{\text{s}}) \max\{\rho-\varepsilon/(\beta\bar{\lambda}), 0\}}{\rho \inf_{j \in \mathbf{N}} (1 - \alpha_j \lambda_j) / (\alpha_j \lambda_j)} \right) \|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2 + 2\mu\eta_{\text{s}}\|\mathbf{z}^* - \mathbf{x}\|^2 \right)}{2\Sigma_k}. \end{aligned}$$

3. Ergodic convergence of PRS: Let \mathbf{z}^* be a fixed point of T_{PRS} . Suppose that \mathbf{f} (respectively \mathbf{g}) is L -Lipschitz, let $\mathbf{x}^k := \mathbf{x}_{\mathbf{g}}^k$ (respectively $\mathbf{x}^k := \mathbf{x}_{\mathbf{f}}^k$), and let $\hat{w} = w$ (respectively $\hat{w} = 1 - w$). Then in Algorithm 4, we have

$$\mathcal{G}^{\text{pre}}(\bar{\mathbf{x}}^k, \bar{\mathbf{x}}^k, \bar{\mathbf{x}}^k; \mathbf{x}) \leq \frac{\|\mathbf{z}^0 - \mathbf{x}\|_U^2 + 4(\gamma/\sqrt{\rho})(L + |\hat{w}|\|\mathbf{S}\|\|\mathbf{x}\|)\|\mathbf{z}^0 - \mathbf{z}^*\|_U}{2\Sigma_k}.$$

4. Ergodic convergence of FBF: Let $\mathbf{z}^* \in \text{zer}(\partial\mathbf{f} + \nabla\mathbf{g} + \mathbf{S})$. Then in Algorithm 5, we have

$$\mathcal{G}^{\text{pre}}(\bar{\mathbf{x}}_{\mathbf{f}}^k, \bar{\mathbf{x}}_{\mathbf{f}}^k, \bar{\mathbf{x}}_{\mathbf{f}}^k; \mathbf{x}) \leq \frac{\|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + 2\eta_{\text{p}}\eta_{\text{s}}\|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2 + 2\mu\eta_{\text{s}}\|\mathbf{z}^* - \mathbf{x}\|^2}{2\Sigma_k}.$$

Proof. Fix $k \in \mathbf{N}$. For any sequence of points $(\mathbf{z}^i)_{i \in \mathbf{N}} \subseteq \mathbf{H}$ and any point $\mathbf{z}^* \in \mathbf{H}$ such that $\|\mathbf{z}^{i+1} - \mathbf{z}^*\|_{U_{i+1}}^2 \leq (1 + \eta_i)\|\mathbf{z}^i - \mathbf{z}^*\|_{U_i}^2$ for all $i \in \mathbf{N}$, we have $\|\mathbf{z}^i - \mathbf{z}^*\|_{U_i}^2 \leq (\prod_{j=0}^{i-1} (1 + \eta_j))\|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2$. Therefore, by the convexity of $\|\cdot\|_{U_i}^2$ for all $i \in \mathbf{N}$, and by the inequality $-\|\mathbf{x}\|_{U_i} \leq -(1/(1 + \eta_i))\|\mathbf{x}\|_{U_{i+1}}$ for all $\mathbf{x} \in \mathbf{H}$ and $i \in \mathbf{N}$, we have

$$\begin{aligned} & \sum_{i=0}^k (\|\mathbf{z}^i - \mathbf{x}\|_{U_i}^2 - \|\mathbf{z}^{i+1} - \mathbf{x}\|_{U_{i+1}}^2) \\ & \leq \|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + \sum_{i=0}^k (\|\mathbf{z}^{i+1} - \mathbf{x}\|_{U_{i+1}}^2 - \|\mathbf{z}^{i+1} - \mathbf{x}\|_{U_i}^2) \\ & \leq \|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + \sum_{i=0}^k \frac{\eta_i}{1 + \eta_i} \|\mathbf{z}^{i+1} - \mathbf{x}\|_{U_{i+1}}^2 \\ & \leq \|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + 2 \sum_{i=0}^k \eta_i (\|\mathbf{z}^{i+1} - \mathbf{z}^*\|_{U_{i+1}}^2 + \|\mathbf{z}^* - \mathbf{x}\|_{U_{i+1}}^2) \\ & \leq \|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + \left(2 \left(\prod_{i=0}^{\infty} (1 + \eta_i)\right) \sum_{i=0}^{\infty} \eta_i\right) \|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2 + 2\mu \left(\sum_{i=0}^{\infty} \eta_i\right) \|\mathbf{z}^* - \mathbf{x}\|^2 \\ & = \|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + 2\eta_{\text{p}}\eta_{\text{s}}\|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2 + 2\mu\eta_{\text{s}}\|\mathbf{z}^* - \mathbf{x}\|^2. \end{aligned} \quad (3.2)$$

We will use Equation (3.2) to produce bounds for all of the variable metric methods.

Part 1: This follows from the Jensen's inequality, Proposition 2.8 ($\kappa_u^i = (1 - 2/\lambda_i)\|\mathbf{z}^{i+1} - \mathbf{z}^i\|_{U_i}^2 \leq 0$), and the fundamental inequality (\mathcal{G}^{pre} does not depend on its second input):

$$\begin{aligned} \mathcal{G}^{\text{pre}}(\bar{\mathbf{x}}_{\mathbf{f}}^k, \bar{\mathbf{x}}_{\mathbf{f}}^k, \bar{\mathbf{x}}_{\mathbf{f}}^k; \mathbf{x}) & \leq \frac{1}{\Sigma_k} \sum_{i=0}^k \gamma_i \lambda_i \mathcal{G}^{\text{pre}}(\mathbf{x}_{\mathbf{f}}^i, \mathbf{x}_{\mathbf{f}}^i, \mathbf{x}_{\mathbf{f}}^i; \mathbf{x}) \\ & \stackrel{(2.11)}{\leq} \frac{1}{2\Sigma_k} \sum_{i=0}^k (\kappa_u^i + \|\mathbf{z}^i - \mathbf{x}\|_{U_i}^2 - \|\mathbf{z}^{i+1} - \mathbf{x}\|_{U_i}^2) \\ & \stackrel{(3.2)}{\leq} \frac{1}{2\Sigma_k} (\|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + 2\eta_{\text{p}}\eta_{\text{s}}\|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2 + 2\mu\eta_{\text{s}}\|\mathbf{z}^* - \mathbf{x}\|^2). \end{aligned}$$

Part 2: We have the following bound from Proposition 2.3:

$$\sum_{i=0}^k \left(\rho - \frac{\varepsilon}{\beta\lambda_i} \right) \|\mathbf{z}^{i+1} - \mathbf{z}^i\|^2 \leq \frac{\max\{\rho - \varepsilon/(\beta\bar{\lambda}), 0\}}{\rho \inf_{j \in \mathbf{N}} (1 - \alpha_j \lambda_j)/(\alpha_j \lambda_j)} (1 + \eta_{\text{p}}\eta_{\text{s}}) \|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2. \quad (3.3)$$

Thus, the bound follows from Jensen's inequality, Proposition 2.8 ($\kappa_u^i \leq (\rho - \varepsilon/(\beta\lambda_i)) \|\mathbf{z}^{i+1} - \mathbf{z}^i\|^2 + 2\gamma_i\lambda_i\mathbf{g}(\mathbf{x}_g^i) - 2\gamma_i\lambda_i\mathbf{g}(\mathbf{x}_f^i)$), and the fundamental inequality:

$$\begin{aligned}
\mathcal{G}^{\text{pre}}(\bar{\mathbf{x}}_f^k, \bar{\mathbf{x}}_f^k, \bar{\mathbf{x}}_f^k; \mathbf{x}) &\leq \frac{1}{\Sigma_k} \sum_{i=0}^k \gamma_i \lambda_i \mathcal{G}^{\text{pre}}(\mathbf{x}_f^i, \mathbf{x}_g^i, \mathbf{x}_f^i; \mathbf{x}) \\
&= \frac{1}{\Sigma_k} \sum_{i=0}^k (\gamma_i \lambda_i \mathcal{G}^{\text{pre}}(\mathbf{x}_f^i, \mathbf{x}_g^i, \mathbf{x}_f^i; \mathbf{x}) + \gamma_i \lambda_i \mathbf{g}(\mathbf{x}_f^i) - \gamma_i \lambda_i \mathbf{g}(\mathbf{x}_g^i)) \\
&\stackrel{(2.11)}{\leq} \frac{1}{2\Sigma_k} \sum_{i=0}^k (\kappa_u^i + \|\mathbf{z}^i - \mathbf{x}\|_{U_i}^2 - \|\mathbf{z}^{i+1} - \mathbf{x}\|_{U_i}^2 + 2\gamma_i \lambda_i \mathbf{g}(\mathbf{x}_f^i) - 2\gamma_i \lambda_i \mathbf{g}(\mathbf{x}_g^i)) \\
&\stackrel{(3.2)}{\leq} \frac{1}{2\Sigma_k} \left(\sum_{i=0}^k \left(\rho - \frac{\varepsilon}{\beta\lambda_i} \right) \|\mathbf{z}^{i+1} - \mathbf{z}^i\|^2 \right) \\
&\quad + \frac{1}{2\Sigma_k} (\|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + 2\eta_p \eta_s \|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2 + 2\mu \eta_s \|\mathbf{z}^* - \mathbf{x}\|^2) \\
&\stackrel{(3.3)}{\leq} \frac{\left(\|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + \left(2\eta_p \eta_s + \frac{(1+\eta_p \eta_s) \max\{\rho - \varepsilon/(\beta\bar{\lambda}), 0\}}{\rho \inf_{j \in \mathbf{N}} (1 - \alpha_j \lambda_j)/(\alpha_j \lambda_j)} \right) \|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2 + 2\mu \eta_s \|\mathbf{z}^* - \mathbf{x}\|^2 \right)}{2\Sigma_k}.
\end{aligned}$$

Part 3: We prove the result when \mathbf{f} is Lipschitz; the other case is symmetric. This follows from the Jensen's inequality, Proposition 2.8 ($\kappa_u^i = (1 - 2/\lambda_i) \|\mathbf{z}^{i+1} - \mathbf{z}^i\|_U^2 \leq 0$), the fundamental inequality, and the identity $\bar{\mathbf{x}}_g^k - \bar{\mathbf{x}}_s^k = w(\bar{\mathbf{x}}_g^k - \bar{\mathbf{x}}_f^k)$ (follows by averaging identities found in Part 3 of Lemma 2.1):

$$\begin{aligned}
\mathcal{G}^{\text{pre}}(\bar{\mathbf{x}}_g^k, \bar{\mathbf{x}}_g^k, \bar{\mathbf{x}}_g^k; \mathbf{x}) &= \mathcal{G}^{\text{pre}}(\bar{\mathbf{x}}_f^k, \bar{\mathbf{x}}_g^k, \bar{\mathbf{x}}_s^k; \mathbf{x}) + \mathbf{f}(\bar{\mathbf{x}}_g^k) - \mathbf{f}(\bar{\mathbf{x}}_f^k) + \langle \mathbf{S}(\bar{\mathbf{x}}_g^k - \bar{\mathbf{x}}_s^k), -\mathbf{x} \rangle \\
&\leq \frac{1}{\Sigma_k} \sum_{i=0}^k \gamma_i \lambda_i \mathcal{G}^{\text{pre}}(\mathbf{x}_f^i, \mathbf{x}_g^i, \mathbf{x}_s^i; \mathbf{x}) \\
&\quad + \mathbf{f}(\bar{\mathbf{x}}_g^k) - \mathbf{f}(\bar{\mathbf{x}}_f^k) + \langle \mathbf{S}(\bar{\mathbf{x}}_g^k - \bar{\mathbf{x}}_s^k), -\mathbf{x} \rangle \\
&\stackrel{(2.11)}{\leq} \frac{1}{2\Sigma_k} \sum_{i=0}^k (\kappa_u^i + \|\mathbf{z}^i - \mathbf{x}\|_U^2 - \|\mathbf{z}^{i+1} - \mathbf{x}\|_U^2) \\
&\quad + L \|\bar{\mathbf{x}}_g^k - \bar{\mathbf{x}}_f^k\| + \|\mathbf{S}\| \|\bar{\mathbf{x}}_g^k - \bar{\mathbf{x}}_s^k\| \|\mathbf{x}\| \\
&\stackrel{(3.1)}{\leq} \frac{\|\mathbf{z}^0 - \mathbf{x}\|_U^2 + 4(\gamma/\sqrt{\rho})(L + |w| \|\mathbf{S}\| \|\mathbf{x}\|) \|\mathbf{z}^0 - \mathbf{z}^*\|_U}{2\Sigma_k}.
\end{aligned}$$

Part 4: This follows from the Jensen's inequality, Proposition 2.8 ($\kappa_u^i \leq 0$), and the fundamental inequality:

$$\begin{aligned}
\mathcal{G}^{\text{pre}}(\bar{\mathbf{x}}_f^k, \bar{\mathbf{x}}_f^k, \bar{\mathbf{x}}_f^k; \mathbf{x}) &\leq \frac{1}{\Sigma_k} \sum_{i=0}^k \gamma_i \lambda_i \mathcal{G}^{\text{pre}}(\mathbf{x}_f^i, \mathbf{x}_f^i, \mathbf{x}_f^i; \mathbf{x}) \\
&\stackrel{(2.11)}{\leq} \frac{1}{2\Sigma_k} \sum_{i=0}^k (\kappa_u^i + \|\mathbf{z}^i - \mathbf{x}\|_{U_i}^2 - \|\mathbf{z}^{i+1} - \mathbf{x}\|_{U_i}^2) \\
&\stackrel{(3.2)}{\leq} \frac{1}{2\Sigma_k} (\|\mathbf{z}^0 - \mathbf{x}\|_{U_0}^2 + 2\eta_p \eta_s \|\mathbf{z}^0 - \mathbf{z}^*\|_{U_0}^2 + 2\mu \eta_s \|\mathbf{z}^* - \mathbf{x}\|^2). \quad \square
\end{aligned}$$

REMARK 2. In general, the $O(1/(k+1))$ convergence rates in Theorem 3.2 are the best PPA, FBS, and PRS obtain for $(\bar{\mathbf{x}}_f^j)_{j \in \mathbf{N}}$ and $(\bar{\mathbf{x}}_g^j)_{j \in \mathbf{N}}$ [24, Proposition 8].

4. Nonergodic convergence. In this section we deduce nonergodic convergence rates for PPA, FBS and PRS under the following assumption:

ASSUMPTION 4. *For all nonergodic convergence results, we assume $(U_j)_{j \in \mathbf{N}}$ and $(\gamma_j)_{j \in \mathbf{N}}$ are constant sequences.*

For PPA, FBS, and PRS, Theorem 4.2 (below) produces a natural sequence $(\mathbf{x}^j)_{j \in \mathbf{N}}$ such that for all bounded subsets $D \subseteq \mathbf{H}$, we have

$$\sup_{\mathbf{x} \in D} \mathcal{G}^{\text{pre}}(\mathbf{x}^k, \mathbf{x}^k, \mathbf{x}^k; \mathbf{x}) = o\left(\frac{1 + \sup_{\mathbf{x} \in D} \|\mathbf{x}\|_U}{\sqrt{k+1}}\right).$$

To the best of our knowledge, the rate of convergence for the nonergodic primal-dual gap generated by the class of algorithms we study has never appeared in the literature.

Nonergodic iterates tend to share structural properties, such as sparsity or low rank, with the solution of the problem. In some cases, the ergodic iterates generated in Section 3 “average out” structural properties of the nonergodic iterates. Thus, although the ergodic iterates may be “closer” to the solution, they are often poorer partial solutions than the nonergodic iterates. The results of this section provide worst-case theoretical guarantees on the quality of the nonergodic iterates in order to justify their use in practical applications.

In our analysis, we use the following result (see also [23] for similar little- o and big- O convergence rates):

THEOREM 4.1 ([24, Theorem 1]). *Let $\alpha \in (0, 1)$, let $\rho > 0$, let $U \in \mathcal{S}_\rho(\mathbf{H})$, and let $(\lambda_j)_{j \in \mathbf{N}} \subseteq (0, 1/\alpha)$. Suppose that $T : \mathbf{H} \rightarrow \mathbf{H}$ is an α -averaged operator in the norm $\|\cdot\|_U$. Let \mathbf{z}^* be a fixed point of T , let $\mathbf{z}^0 \in \mathbf{H}$, let $\tau_k := (1 - \alpha\lambda_k)\lambda_k/\alpha$ for all $k \in \mathbf{N}$, suppose that $\underline{\tau} := \inf_{j \in \mathbf{N}} \tau_j > 0$, and suppose that $(\mathbf{z}^j)_{j \in \mathbf{N}}$ is generated by the following iteration: for all $k \in \mathbf{N}$, let*

$$\mathbf{z}^{k+1} := T_{\lambda_k}(\mathbf{z}^k). \quad (4.1)$$

Then for all $k \in \mathbf{N}$, we have

$$\|T\mathbf{z}^k - \mathbf{z}^k\|_U^2 \leq \frac{\|\mathbf{z}^0 - \mathbf{z}^*\|_U^2}{\underline{\tau}(k+1)} \quad \text{and} \quad \|T\mathbf{z}^k - \mathbf{z}^k\|_U^2 = o\left(\frac{1}{k+1}\right). \quad (4.2)$$

Throughout this section, T will always denote an α -averaged mapping in the norm $\|\cdot\|_U$. Recall that for $\lambda \in (0, 1/\alpha)$, T_λ is $\alpha\lambda$ -averaged (see Proposition 1.2), so

$$\|T_\lambda \mathbf{z}^k - \mathbf{z}^*\|_U^2 \stackrel{(1.8)}{\leq} \|\mathbf{z}^k - \mathbf{z}^*\|_U^2 - \frac{1 - \alpha\lambda}{\alpha\lambda} \|T_\lambda \mathbf{z}^k - \mathbf{z}^k\|_U^2 \quad (4.3)$$

for all $k \in \mathbf{N}$, and any fixed-point \mathbf{z}^* of T . Note that Equation (4.3) also holds when $\alpha\lambda = 1$ (see Proposition 1.2). Equation (4.3) shows that $T_\lambda \mathbf{z}^k$ is at least as close to \mathbf{z}^* as \mathbf{z}^k is. This fact will be useful in the proof of Theorem 4.2 below.

In the following theorem, we will deduce little- o and big- O convergence rates. Because the pre-primal-dual gap can be negative, we slightly abuse notation: given a point $\mathbf{x} \in \mathbf{H}$, a (not necessarily positive) sequence $(a_j)_{j \in \mathbf{N}}$ satisfies $a_k = o((1 + \|\mathbf{x}\|_U)/\sqrt{k+1})$ provided that there exists a nonnegative sequence $(b_j)_{j \in \mathbf{N}}$ such that $b_k = o((1 + \|\mathbf{x}\|_U)/\sqrt{k+1})$ and $a_k = O(b_k)$. Note that we do not measure $|a_k|$ because our only goal is to ensure that the sequence $(a_j)_{j \in \mathbf{N}}$ is eventually nonpositive.

THEOREM 4.2. *Suppose that Assumption 4 holds, let $U \in \mathcal{S}_\rho(\mathbf{H})$ denote the common metric inducing map, and let $\gamma \in \mathbf{R}_{++}$ denote the common stepsize parameter. Then each method is a special case of Iteration (4.1). For each method, assume that $\underline{\tau} > 0$ (See Theorem 4.1). Then for all $k \in \mathbf{N}$ and all $\mathbf{x} \in \mathbf{H}$, the following hold:*

1. **Nonergodic convergence of PPA:** Let $\mathbf{z}^* \in \text{zer}(\partial \mathbf{f} + \mathbf{S})$. Then in Algorithm 2, we have $\alpha = 1/2$ and $T = J_{U^{-1}(\partial \mathbf{f} + \mathbf{S})}$,

$$\mathcal{G}^{\text{pre}}(\mathbf{x}_f^k, \mathbf{x}_f^k, \mathbf{x}_f^k; \mathbf{x}) \leq \frac{(\|\mathbf{z}^0 - \mathbf{z}^*\|_U + \|\mathbf{z}^* - \mathbf{x}\|_U) \|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\gamma \sqrt{\mathcal{I}(k+1)}},$$

and $\mathcal{G}^{\text{pre}}(\mathbf{x}_f^k, \mathbf{x}_f^k, \mathbf{x}_f^k; \mathbf{x}) = o((1 + \|\mathbf{x}\|_U)/\sqrt{k+1})$.

2. **Nonergodic convergence of FBS:** Let $\mathbf{z}^* \in \text{zer}(\partial \mathbf{f} + \nabla \mathbf{g} + \mathbf{S})$. Then in Algorithm 3, we have $\alpha = \alpha_{\gamma, \rho}$ (Equation (2.3)) and $T = T_{\text{FBS}}^{U, \gamma}$ (Equation (2.2)),

$$\mathcal{G}^{\text{pre}}(\mathbf{x}_f^k, \mathbf{x}_f^k, \mathbf{x}_f^k; \mathbf{x}) \leq \frac{(\|\mathbf{z}^0 - \mathbf{z}^*\|_U + \|\mathbf{z}^* - \mathbf{x}\|_U) \|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\gamma \sqrt{\mathcal{I}(k+1)}},$$

and $\mathcal{G}^{\text{pre}}(\mathbf{x}_f^k, \mathbf{x}_f^k, \mathbf{x}_f^k; \mathbf{x}) = o((1 + \|\mathbf{x}\|_U)/\sqrt{k+1})$.

3. **Nonergodic convergence of PRS:** Let \mathbf{z}^* be a fixed point of T_{PRS} (Equation (2.4)). Then in Algorithm 4, we have $\alpha = 1/2$ and $T = (T_{\text{PRS}})_{1/2}$ (Equation (2.5)). In addition, suppose that \mathbf{f} (respectively \mathbf{g}) is L -Lipschitz, let $\mathbf{x}^k := \mathbf{x}_g^k$ (respectively $\mathbf{x}^k := \mathbf{x}_f^k$), and let $\hat{w} = w$ (respectively $\hat{w} = 1 - w$). Then

$$\mathcal{G}^{\text{pre}}(\mathbf{x}^k, \mathbf{x}^k, \mathbf{x}^k; \mathbf{x}) \leq \frac{(\|\mathbf{z}^0 - \mathbf{z}^*\|_U + \|\mathbf{z}^* - \mathbf{x}\|_U + (\gamma/\sqrt{\rho})(L + \hat{w}\|\mathbf{S}\|\|\mathbf{x}\|)) \|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\gamma \sqrt{\mathcal{I}(k+1)}},$$

and $\mathcal{G}^{\text{pre}}(\mathbf{x}^k, \mathbf{x}^k, \mathbf{x}^k; \mathbf{x}) = o((1 + \|\mathbf{x}\|_U)/\sqrt{k+1})$.

Proof. Fix $k \in \mathbf{N}$. In all of the following proofs, we will bound the pre-primal-dual gap by a quantity involving $\|T\mathbf{z}^k - \mathbf{z}^k\|_U$. Then the big- O and little- o convergence rates follow directly from Theorem 4.1. In addition, we will use Equation (4.3) and the independence of $\mathbf{x}_f^k, \mathbf{x}_g^k$, and \mathbf{x}_S^k from λ_k to tighten our upper bounds. To this end, we will denote $\mathbf{z}_\lambda := T_\lambda(\mathbf{z}^k)$ (see Equation (1.3)) and let $C = (0, 1/\alpha]$ where α is averagedness coefficient of T . Note that T_λ is nonexpansive for all $\lambda \in C$ (see Part 3 of Proposition 1.2). Also note that for $\lambda \in C$, we have $(1/\lambda)(\mathbf{z}_\lambda - \mathbf{z}^k) = T\mathbf{z}^k - \mathbf{z}^k$ and $\|\mathbf{z}_\lambda - \mathbf{z}^*\|_U \leq \|\mathbf{z}^k - \mathbf{z}^*\|_U \leq \|\mathbf{z}^0 - \mathbf{z}^*\|_U$ by Equation (4.3) and the monotonicity of $(\|\mathbf{z}^j - \mathbf{z}^*\|_U)_{j \in \mathbf{N}}$ (Proposition 2.3). Thus, $\|\mathbf{z}_\lambda - \mathbf{x}\|_U \leq \|\mathbf{z}^0 - \mathbf{z}^*\|_U + \|\mathbf{z}^* - \mathbf{x}\|_U$. Therefore, for all $\lambda \in (0, 1/\alpha]$, we have

$$\frac{\langle \mathbf{z}^k - \mathbf{z}_\lambda, \mathbf{z}_\lambda - \mathbf{x} \rangle_U}{\lambda} \leq \|T\mathbf{z}^k - \mathbf{z}^k\|_U \|\mathbf{z}_\lambda - \mathbf{x}\|_U \stackrel{(4.2)}{\leq} \frac{(\|\mathbf{z}^0 - \mathbf{z}^*\|_U + \|\mathbf{z}^* - \mathbf{x}\|_U) \|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\sqrt{\mathcal{I}(k+1)}}. \quad (4.4)$$

Note that the upper key term identities (Proposition 2.8) and the fundamental inequality (Proposition 2.6) continue to hold when \mathbf{z}^{k+1} is replaced by \mathbf{z}_λ . Thus, in each of the cases below, we will minimize the fundamental inequality over all $\lambda \in C$.

Part 1: Proposition 2.8 shows that $\kappa_u^k(\lambda) = (1 - 2/\lambda) \|\mathbf{z}_\lambda - \mathbf{z}^k\|_U^2$. Thus, the fundamental inequality, the cosine rule, and the identity $C = (0, 2]$ show (\mathcal{G}^{pre} does not depend on its second input)

$$\begin{aligned} \mathcal{G}^{\text{pre}}(\mathbf{x}_f^k, \mathbf{x}_f^k, \mathbf{x}_f^k; \mathbf{x}) &\leq \inf_{\lambda \in C} \frac{1}{2\gamma\lambda} \left(\left(1 - \frac{2}{\lambda}\right) \|\mathbf{z}_\lambda - \mathbf{z}^k\|_U^2 + \|\mathbf{z}^k - \mathbf{x}\|_U^2 - \|\mathbf{z}_\lambda - \mathbf{x}\|_U^2 \right) \\ &\stackrel{(1.7)}{=} \inf_{\lambda \in C} \frac{1}{2\gamma\lambda} \left(2\langle \mathbf{z}^k - \mathbf{z}_\lambda, \mathbf{z}_\lambda - \mathbf{x} \rangle_U + 2\left(1 - \frac{1}{\lambda}\right) \|\mathbf{z}_\lambda - \mathbf{z}^k\|_U^2 \right) \\ &\leq \frac{1}{\gamma} \langle \mathbf{z}^k - \mathbf{z}_1, \mathbf{z}_1 - \mathbf{x} \rangle_U \stackrel{(4.4)}{\leq} \frac{(\|\mathbf{z}^0 - \mathbf{z}^*\|_U + \|\mathbf{z}^* - \mathbf{x}\|_U) \|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\gamma \sqrt{\mathcal{I}(k+1)}}. \end{aligned}$$

Part 2: First choose $\tilde{\lambda} \in C$ small enough that $\rho + \mu - \varepsilon/(\beta\tilde{\lambda}) \leq 0$. Now recall that Proposition 2.8 proves the following inequality: $\kappa_u^k(\lambda) \leq (\rho - \varepsilon/(\beta\lambda)) \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + 2\gamma\lambda\mathbf{g}(\mathbf{x}_f^k) - 2\gamma\lambda\mathbf{g}(\mathbf{x}_f^k)$. Thus, the fundamental inequality, the cosine rule, and the identity $C = (0, 1/\alpha]$ show

$$\begin{aligned}
\mathcal{G}^{\text{pre}}(\mathbf{x}_f^k, \mathbf{x}_f^k, \mathbf{x}_f^k; \mathbf{x}) &= \mathcal{G}^{\text{pre}}(\mathbf{x}_f^k, \mathbf{x}_g^k, \mathbf{x}_f^k; \mathbf{x}) + \mathbf{g}(\mathbf{x}_f^k) - \mathbf{g}(\mathbf{x}_g^k) \\
&\leq \inf_{\lambda \in C} \frac{1}{2\gamma\lambda} (2\gamma\lambda\mathbf{g}(\mathbf{x}_f^k) - 2\gamma\lambda\mathbf{g}(\mathbf{x}_g^k) + \kappa_u^k(\lambda) + \|\mathbf{z}^k - \mathbf{x}\|_U^2 - \|\mathbf{z}_\lambda - \mathbf{x}\|_U^2) \\
&\leq \inf_{\lambda \in C} \frac{1}{2\gamma\lambda} \left(\left(\rho - \frac{\varepsilon}{\beta\lambda} \right) \|\mathbf{z}_\lambda - \mathbf{z}^k\|^2 + \|\mathbf{z}^k - \mathbf{x}\|_U^2 - \|\mathbf{z}_\lambda - \mathbf{x}\|_U^2 \right) \\
&\stackrel{(1.7)}{=} \inf_{\lambda \in C} \frac{1}{2\gamma\lambda} \left(2\langle \mathbf{z}^k - \mathbf{z}_\lambda, \mathbf{z}_\lambda - \mathbf{x} \rangle_U + \|\mathbf{z}_\lambda - \mathbf{z}^k\|_U^2 + \left(\rho - \frac{\varepsilon}{\beta\lambda} \right) \|\mathbf{z}_\lambda - \mathbf{z}^k\|^2 \right) \\
&\leq \inf_{\lambda \in C} \frac{1}{2\gamma\lambda} \left(2\langle \mathbf{z}^k - \mathbf{z}_\lambda, \mathbf{z}_\lambda - \mathbf{x} \rangle_U + \left((\rho + \mu) - \frac{\varepsilon}{\beta\lambda} \right) \|\mathbf{z}_\lambda - \mathbf{z}^k\|^2 \right) \\
&\leq \frac{1}{\gamma\tilde{\lambda}} \langle \mathbf{z}^k - \mathbf{z}_{\tilde{\lambda}}, \mathbf{z}_{\tilde{\lambda}} - \mathbf{x} \rangle_U \stackrel{(4.4)}{\leq} \frac{(\|\mathbf{z}^0 - \mathbf{z}^*\|_U + \|\mathbf{z}^* - \mathbf{x}\|_U) \|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\gamma\sqrt{\mathcal{I}(k+1)}}.
\end{aligned}$$

Part 3: We prove the result in the case that \mathbf{f} is Lipschitz because the other case is symmetric. Proposition 2.8 proves the following identity: $\kappa_u^k(\lambda) = (1 - 2/\lambda) \|\mathbf{z}_\lambda - \mathbf{z}^k\|_U^2$. Thus, the fundamental inequality, the cosine rule, and the identities $\mathbf{x}_f^k - \mathbf{x}_g^k = (1/\lambda)(\mathbf{z}_\lambda - \mathbf{z}^k) = T\mathbf{z}^k - \mathbf{z}^k$, $\mathbf{x}_g^k - \mathbf{x}_S^k = w(\mathbf{x}_g^k - \mathbf{x}_f^k)$, and $C = (0, 2]$ show

$$\begin{aligned}
\mathcal{G}^{\text{pre}}(\mathbf{x}_g^k, \mathbf{x}_g^k, \mathbf{x}_g^k; \mathbf{x}) &\leq \mathcal{G}^{\text{pre}}(\mathbf{x}_f^k, \mathbf{x}_g^k, \mathbf{x}_S^k; \mathbf{x}) + \mathbf{f}(\mathbf{x}_g^k) - \mathbf{f}(\mathbf{x}_f^k) + \langle \mathbf{S}(\mathbf{x}_g^k - \mathbf{x}_S^k), -\mathbf{x} \rangle \\
&\leq \inf_{\lambda \in C} \frac{1}{2\gamma\lambda} \left(\left(1 - \frac{2}{\lambda} \right) \|\mathbf{z}_\lambda - \mathbf{z}^k\|_U^2 + \|\mathbf{z}^k - \mathbf{x}\|_U^2 - \|\mathbf{z}_\lambda - \mathbf{x}\|_U^2 \right) \\
&\quad + L\|\mathbf{x}_g^k - \mathbf{x}_f^k\| + |w|\|\mathbf{S}\|\|\mathbf{x}_g^k - \mathbf{x}_f^k\|\|\mathbf{x}\| \\
&\stackrel{(1.7)}{=} \inf_{\lambda \in C} \frac{1}{2\gamma\lambda} \left(2\langle \mathbf{z}^k - \mathbf{z}_\lambda, \mathbf{z}_\lambda - \mathbf{x} \rangle_U + 2 \left(1 - \frac{1}{\lambda} \right) \|\mathbf{z}_\lambda - \mathbf{z}^k\|_U^2 \right) \\
&\quad + L\|\mathbf{x}_g^k - \mathbf{x}_f^k\| + |w|\|\mathbf{S}\|\|\mathbf{x}_g^k - \mathbf{x}_f^k\|\|\mathbf{x}\| \\
&\leq \frac{1}{\gamma} \langle \mathbf{z}^k - \mathbf{z}_1, \mathbf{z}_1 - \mathbf{x} \rangle_U + L\|\mathbf{x}_g^k - \mathbf{x}_f^k\| + |w|\|\mathbf{S}\|\|\mathbf{x}_g^k - \mathbf{x}_f^k\|\|\mathbf{x}\| \\
&\stackrel{(4.4)}{\leq} \frac{(\|\mathbf{z}^0 - \mathbf{z}^*\|_U + \|\mathbf{z}^* - \mathbf{x}\|_U) \|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\gamma\sqrt{\mathcal{I}(k+1)}} \\
&\stackrel{(4.2)}{+} \frac{(L + |w|\|\mathbf{S}\|\|\mathbf{x}\|)\|\mathbf{z}^0 - \mathbf{z}^*\|_U}{\sqrt{\rho\mathcal{I}(k+1)}}. \quad \square
\end{aligned}$$

REMARK 3. Note that we can immediately strengthen the convergence result for PRS in Theorems 4.2 and 3.2. Indeed, we only need to assume that \mathbf{f} or \mathbf{g} is Lipschitz on the closed ball $B_U(\mathbf{x}^*; \|\mathbf{z}^0 - \mathbf{z}^*\|_U)$ (where $\mathbf{x}^* = J_{\gamma U^{-1}(\partial\mathbf{g} + (1-w)\mathbf{S})}(\mathbf{z}^*)$) of radius $\|\mathbf{z}^0 - \mathbf{z}^*\|_U$ (under the metric $\|\cdot\|_U$) because for all $k \in \mathbb{N}$,

$$\begin{aligned}
\|\mathbf{x}_g^k - \mathbf{x}^*\|_U &= \|J_{\gamma U^{-1}(\partial\mathbf{g} + (1-w)\mathbf{S})}(\mathbf{z}^k) - J_{\gamma U^{-1}(\partial\mathbf{g} + (1-w)\mathbf{S})}(\mathbf{z}^*)\|_U \leq \|\mathbf{z}^k - \mathbf{z}^*\|_U \\
&\leq \|\mathbf{z}^0 - \mathbf{z}^*\|_U,
\end{aligned}$$

and by a similar derivation, $\|\mathbf{x}_f^k - \mathbf{x}^*\|_U \leq \|\mathbf{z}^0 - \mathbf{z}^*\|_U$. Thus, the sequences lie in the ball: $(\mathbf{x}_f^j)_{j \in \mathbb{N}}, (\mathbf{x}_g^j)_{j \in \mathbb{N}} \subseteq B_U(\mathbf{x}^*, \|\mathbf{z}^0 - \mathbf{z}^*\|_U)$. We also have $(\bar{\mathbf{x}}_f^j)_{j \in \mathbb{N}}, (\bar{\mathbf{x}}_g^j)_{j \in \mathbb{N}} \subseteq$

$\overline{B_U(\mathbf{x}^*, \|\mathbf{z}^0 - \mathbf{z}^*\|_U)}$ by the convexity of the ball. See [2, Proposition 8.28] for conditions that ensure Lipschitz continuity of convex functions on balls.

REMARK 4. In general, the $o(1/\sqrt{k+1})$ convergence rates in Theorem 4.2 are the best PRS can obtain for $(\mathbf{x}_g^j)_{j \in \mathbb{N}}$ [24, Theorem 11].

REMARK 5. In general, it is infeasible to take the supremum over the last component of \mathcal{G}^{pre} as in Equation (2.9). Thus, in practice we cannot use the pre-primal-dual gap to measure convergence. However, Theorem 4.2 bounds the pre-primal-dual gap at the k -th iteration by a multiple of the expression $\|T\mathbf{z}^k - \mathbf{z}^k\| \|\mathbf{x}\|$. Thus, if the supremum in Equation (2.9) can be restricted to a bounded set D , then $\|T\mathbf{z}^k - \mathbf{z}^k\| \sup_{\mathbf{x} \in D} \|\mathbf{x}\|$ can be used as a proxy for the size of the pre-primal-dual gap. See section 5.1 for examples of such sets D .

5. Applications. In this section we will show that the four algorithms from Section 2.2 are capable of solving highly structured optimization problems:

PROBLEM 2 (Model problem). Let \mathcal{H}_0 be a Hilbert space, and let $f, g : \Gamma_0(\mathcal{H}_0)$. Let $n \in \mathbb{N} \setminus \{0\}$, and for $i = 1, \dots, n$, let \mathcal{H}_i be a Hilbert space, let $h_i, l_i \in \Gamma_0(\mathcal{H}_i)$, suppose that $h_i \square l_i \in \Gamma_0(\mathcal{H}_i)$, and let $B_i : \mathcal{H}_0 \rightarrow \mathcal{H}_i$ be a bounded linear map. Finally, let $\mathbf{B} : \mathcal{H}_0 \rightarrow \prod_{i=1}^n \mathcal{H}_i$ be the map $x \mapsto (B_1x, \dots, B_nx)$. Then our model problem is as follows:

$$\underset{x \in \mathcal{H}_0}{\text{minimize}} f(x) + g(x) + \sum_{i=1}^n (h_i \square l_i)(B_i x). \quad (5.1)$$

In addition, the dual problem is to

$$\underset{\mathbf{y} \in \prod_{i=1}^n \mathcal{H}_i}{\text{minimize}} (f^* \square g^*)(-\mathbf{B}^* \mathbf{y}) + \sum_{i=1}^n (h_i^* + l_i^*)(y_i).$$

All of the algorithms we consider take full advantage of the structure of the infimal convolution in Problem 2. We note that infimal convolutions are not widespread in applications. Generally, for $i \in \{1, \dots, n\}$, we think of $h_i \square l_i$ as a regularization of h_i by l_i , or vice versa. Indeed, under mild conditions, the smoothness of at least one of h_i and l_i implies the smoothness of the infimal convolution [2, Section 18.3]. When l_i or h_i is chosen properly, this operation is sometimes called *dual-smoothing* [34]. Finally, we note that we can remove the infimal convolution operation from Problem 2 by setting $l_i = \iota_{\{0\}}$ because $h_i \square l_i = h_i$ for all $i = 1, \dots, n$. The interested reader should consult [2, Proposition 12.14 and Proposition 15.7] for conditions that guarantee that $h_i \square l_i \in \Gamma_0(\mathcal{H}_i)$.

We assume the existence of a specific type of solution of Problem 2.

ASSUMPTION 5. We assume that there exists

$$x^* \in \text{zer} \left(\partial f + \partial g + \sum_{i=1}^n B_i^* (\partial h_i \square \partial l_i)(B_i(\cdot)) \right).$$

See [19, Proposition 4.3] for conditions that guarantee the existence of x^* . In general, the containment

$$\text{zer} \left(\partial f + \partial g + \sum_{i=1}^n B_i^* (\partial h_i \square \partial l_i)(B_i(\cdot)) \right) \subseteq \text{zer} \left(\partial \left(f + g + \sum_{i=1}^n (h_i \square l_i)(B_i(\cdot)) \right) \right)$$

always holds, but the sets may not be equal. Nevertheless, this assumption is standard.

We now review two possible splittings of Problem 2. Both splittings will be designated by a “level.” The level is an indication of the number of extra dual variables that are introduced into the problem. Introducing more dual variables makes the problem further separable, and, hence, further parallelizable, but it also increases the memory footprint of the algorithm. It is unclear whether the number of dual variables affects the practical convergence speed of the algorithm in a negative way.

The following proposition is a simple exercise in duality, so we omit the proof.

PROPOSITION 5.1 (Level 1 optimality conditions). *Let $\mathbf{H} = \prod_{i=0}^n \mathcal{H}_i$, and denote an arbitrary point $\mathbf{x} \in \mathbf{H}$ by $\mathbf{x} = (x, y_1, \dots, y_n) = (x, \mathbf{y})$. For all $\mathbf{x} \in \mathbf{H}$, let $\mathbf{f}(\mathbf{x}) := f(x) + \sum_{i=1}^n h_i^*(y_i)$, let $\mathbf{g}(\mathbf{x}) := g(x) + \sum_{i=1}^n l_i^*(y_i)$, and let $\mathbf{S} : \mathbf{H} \rightarrow \mathbf{H}$ be the skew map $(x, \mathbf{y}) \mapsto (\mathbf{B}^* \mathbf{y}, -\mathbf{B}x)$. Then a point $x^* \in \mathcal{H}_0$ satisfies*

$$0 \in \partial f(x^*) + \partial g(x^*) + \sum_{i=1}^n B_i^* (\partial h_i \square \partial l_i) (B_i x^*) \quad (5.2)$$

if, and only if, there is a vector $\mathbf{y}^ \in \prod_{i=1}^n \mathcal{H}_i$ such that*

$$0 \in \partial \mathbf{f}(x^*, \mathbf{y}^*) + \partial \mathbf{g}(x^*, \mathbf{y}^*) + \mathbf{S}(x^*, \mathbf{y}^*). \quad (5.3)$$

Notice that the subdifferential operators $\partial \mathbf{f}$ and $\partial \mathbf{g}$ in Equation (5.3) are completely separable in the variables of the product space \mathbf{H} . Thus, evaluating the proximity operators of \mathbf{f} and \mathbf{g} can be quite simple. However, the resolvent $J_{\partial \mathbf{f} + \mathbf{S}}$ is not necessarily simple to evaluate. This difficulty motivates the introduction of new metrics on \mathbf{H} that simplify the resolvent computation (Section 5.2).

Whenever the functions g and l_i^* are Lipschitz differentiable for $i \in \{1, \dots, n\}$ (or equivalently, l_i is strongly convex [2, Theorem 18.15]) we can apply FBS or FBF (Algorithms 3 and 5) to the splitting in Proposition 5.1. For nonsmooth g and l_i^* , we can apply the PRS algorithm.

The proof of the following proposition is similar to Proposition 5.1, so we omit it. The proposition is most useful in the case that g or l_i^* are not differentiable for some $i \in \{1, \dots, n\}$.

PROPOSITION 5.2 (Level 2 optimality conditions). *Let $\mathbf{H} = \mathcal{H}_0 \times (\prod_{i=1}^n \mathcal{H}_i)^2$, and denote an arbitrary $\mathbf{x} \in \mathbf{H}$ by $\mathbf{x} = (x, y_1, \dots, y_n, v_1, \dots, v_n) = (x, \mathbf{y}, \mathbf{v})$. For all $\mathbf{x} \in \mathbf{H}$, let $\mathbf{f}(\mathbf{x}) := f(x) + \sum_{i=1}^n (h_i^*(y_i) + l_i(v_i))$, let $\mathbf{g}(\mathbf{x}) := g(x)$, and let $\mathbf{S} : \mathbf{H} \rightarrow \mathbf{H}$ be the skew map $(x, \mathbf{y}, \mathbf{v}) \mapsto (\mathbf{B}^* \mathbf{y}, -\mathbf{B}x + \mathbf{v}, -\mathbf{y})$. Then a point $x^* \in \mathcal{H}_0$ satisfies*

$$0 \in \partial f(x^*) + \partial g(x^*) + \sum_{i=1}^n B_i^* (\partial h_i \square \partial l_i) (B_i x^*) \quad (5.4)$$

if, and only if, there is a vector $(\mathbf{y}^, \mathbf{v}^*) \in (\prod_{i=1}^n \mathcal{H}_i)^2$ such that*

$$0 \in \partial \mathbf{f}(x^*, \mathbf{y}^*, \mathbf{v}^*) + \partial \mathbf{g}(x^*, \mathbf{y}^*, \mathbf{v}^*) + \mathbf{S}(x^*, \mathbf{y}^*, \mathbf{v}^*). \quad (5.5)$$

Note that if for some $i \in \{1, \dots, n\}$, l_i is differentiable, we can “assign” it to the function \mathbf{g} instead of “assigning” it to \mathbf{f} . If g is also differentiable, we can apply FBS to the inclusion.

There are many splittings that solve Problem 2. Furthermore, the complexity of Problem 2 can be increased in various ways, e.g., by precomposing each of h_i and l_i with linear operators [3, 10], or by solving systems of such inclusions [17, 8]. We choose to discuss this relatively simple formulation for clarity of exposition.

The next several sections relate the results and notation of the previous sections to the level 1 and 2 splittings.

5.1. Primal-dual gap functions. In this section, we discuss the pre-primal-dual gap function in the context of the level 1 splitting in Proposition 5.1. We give sufficient conditions for the gap function (Definition 2.5) to bound the primal and dual objectives of Problem 2 and show that the pre-primal-dual gap also bounds certain squared norms that arise from the strong convexity and differentiability of the terms of the objective.

In the level 1 splitting, the pre-primal-dual gap has the following form: for all $(x, \mathbf{y}), (x^*, \mathbf{y}^*) \in \mathbf{H}$ (with components defined as in Proposition 5.1), we have

$$\begin{aligned} \mathcal{G}^{\text{pre}}(\mathbf{x}, \mathbf{x}, \mathbf{x}; \mathbf{x}^*) &= f(x) + g(x) - f(x^*) - g(x^*) + \langle x - x^*, \mathbf{B}^* \mathbf{y}^* \rangle \\ &\quad + \sum_{i=1}^n (h_i^*(y_i) + l_i^*(y_i) - h_i^*(y_i^*) - l_i^*(y_i^*)) - \langle \mathbf{B} \mathbf{x}^*, \mathbf{y} - \mathbf{y}^* \rangle, \end{aligned} \quad (5.6)$$

where we used the identity $\langle \mathbf{S} \mathbf{x}, -\mathbf{x}^* \rangle = \langle \mathbf{S} \mathbf{x}, \mathbf{x} - \mathbf{x}^* \rangle$. If \mathbf{x}^* satisfies the inclusion in Proposition 5.1, then

$$-\mathbf{B}^* \mathbf{y}^* \in \partial f(x^*) + \partial g(x^*) \quad \text{and} \quad B_i x^* \in \partial h_i^*(y_i^*) + \partial l_i^*(y_i^*). \quad (5.7)$$

We will now bound several terms that arise from the strong convexity and Lipschitz differentiability of the terms in the objective function.

We follow the convention that every closed, proper, and convex function $F : \mathcal{H}_0 \rightarrow (-\infty, \infty]$ is μ_F -strongly convex and $\tilde{\nabla} F$ is L_F -Lipschitz for some $\mu_F \in \mathbf{R}_+$ and $L_F \in [0, +\infty]$. If F is not differentiable, then we let $L_F = \infty$. In addition, if $L_F < \infty$, then $\tilde{\nabla} F = \nabla F$ is Lipschitz. Note that we allow the $\mu_F = 0$. The following quantity is useful for summarizing the lower bounds that we derive from strong convexity and Lipschitz differentiability: for all $x \in \mathcal{H}_0$ and $y \in \text{dom}(\partial F)$, if

$$S_F(x, y) := \begin{cases} \max \left\{ \frac{\mu_F}{2} \|x - y\|^2, \frac{1}{2L_F} \|\nabla F(x) - \nabla F(y)\|^2 \right\} & \text{if } L_F < \infty; \\ \frac{\mu_F}{2} \|x - y\|^2 & \text{otherwise;} \end{cases} \quad (5.8)$$

then combine [2, Theorem 18.15(iv) and Proposition 16.9] to get

$$F(x) \geq F(y) + \langle x - y, \tilde{\nabla} F(y) \rangle + S_F(x, y). \quad (5.9)$$

We use the analogous notation for f, g and the conjugate functions h_i^*, l_i^* for $i = 1, \dots, n$. Therefore, if we apply the lower bound in Equation (5.9) to each of the functions in Equation (5.6) and use the subgradient identities in Equation (5.7) to cancel inner products, we get

$$\mathcal{G}^{\text{pre}}(\mathbf{x}, \mathbf{x}, \mathbf{x}; \mathbf{x}^*) \geq S_f(x, x^*) + S_g(x, x^*) + \sum_{i=1}^n (S_{h_i^*}(y_i, y_i^*) + S_{l_i^*}(y_i, y_i^*)). \quad (5.10)$$

Equation (5.10) shows that convergence rates for the pre-primal-dual gap function immediately imply the same convergence rates for the $S(\cdot, \cdot)$ functions in Equation (5.8). Note that this lower bound does not require that $\text{dom}(\mathbf{f})$ or $\text{dom}(\mathbf{g})$ are bounded.

The next proposition gives sufficient conditions under which the pre-primal-dual gap bounds the primal and dual objectives. In general, we cannot expect such a bound to hold, unless several terms in the objective are Lipschitz continuous or certain subdifferentials are locally bounded.

PROPOSITION 5.3 (Level 1 gap function bounds). *Let x^* be a minimizer of Problem 2. Assume the notation of Proposition 5.1. Let $D_1 \subseteq \mathcal{H}$ and let $D_2 \subseteq$*

$\prod_{i=1}^n \mathcal{H}_i$ be bounded sets. Then for any sequence of points $((x^j, \mathbf{y}^j))_{j \geq 0} \subseteq \text{dom}(f + g) \times \prod_{i=1}^n \text{dom}(h_i^* + l_i^*)$, the inequality

$$\begin{aligned} & f(x^k) + g(x^k) + \sum_{i=1}^n (h_i \square l_i)(B_i x^k) - \left(f(x^*) + g(x^*) + \sum_{i=1}^n (h_i \square l_i)(B_i x^*) \right) \\ & \leq \sup_{\mathbf{x} \in \{x^*\} \times D_2} \mathcal{G}^{\text{pre}}(\mathbf{x}^k, \mathbf{x}^k, \mathbf{x}^k; \mathbf{x}) \end{aligned}$$

holds for all $k \in \mathbb{N}$ provided either of the following hold:

1. $\text{dom}(h_1^* + l_1^*) \times \cdots \times \text{dom}(h_n^* + l_n^*) \subseteq D_2$;
2. $\partial(h_1 \square l_1)(B_1 x^k) \times \cdots \times \partial(h_n \square l_n)(B_n x^k) \subseteq D_2$.

Similarly, the inequality

$$\begin{aligned} & (f^* \square g^*)(-\mathbf{B}^* \mathbf{y}^k) + \sum_{i=1}^n (h_i^* + l_i^*)(y_i^k) - \left((f^* \square g^*)(-\mathbf{B}^* \mathbf{y}^*) + \sum_{i=1}^n (h_i^* + l_i^*)(B_i y_i^*) \right) \\ & \leq \sup_{\mathbf{x} \in D_1 \times \{\mathbf{y}^*\}} \mathcal{G}^{\text{pre}}(\mathbf{x}^k, \mathbf{x}^k, \mathbf{x}^k; \mathbf{x}) \end{aligned}$$

holds for all $k \in \mathbb{N}$ provided either of the following hold:

1. $\text{dom}(f + g) \subseteq D_1$;
2. $\partial(f^* \square g^*)(-\mathbf{B}^* \mathbf{y}^k) \subseteq D_1$.

Proof. Fix $k \in \mathbb{N}$. We only consider the primal case because the dual case is similar. For all $i \in \{1, \dots, n\}$, the Fenchel-Moreau Theorem [2, Theorem 13.32], the identity $h_i \square l_i = (h_i^* + l_i^*)^*$, and Conditions 1 and 2 show that we can reduce the domain of the following supremum:

$$\begin{aligned} \sum_{i=1}^n (h_i \square l_i)(B_i x^k) &= \sup_{\mathbf{y} \in \mathbf{H}} \left(\langle \mathbf{B} x^k, \mathbf{y} \rangle - \sum_{i=1}^n (h_i^*(y_i) + l_i^*(y_i)) \right) \\ &= \sup_{\mathbf{y} \in D_2} \left(\langle \mathbf{B} x^k, \mathbf{y} \rangle - \sum_{i=1}^n (h_i^*(y_i) + l_i^*(y_i)) \right). \end{aligned}$$

In addition, the Fenchel-Young inequality shows that

$$\sum_{i=1}^n (h_i^*(y_i^k) + l_i^*(y_i^k)) - \langle x^*, \mathbf{B}^* \mathbf{y}^k \rangle \geq - \sum_{i=1}^n (h_i \square l_i)(B_i x^*).$$

Therefore,

$$\begin{aligned} & \sup_{\mathbf{x} \in \{x^*\} \times D_2} \mathcal{G}^{\text{pre}}(\mathbf{x}^k, \mathbf{x}^k, \mathbf{x}^k; \mathbf{x}) \\ &= f(x^k) + g(x^k) - f(x^*) - g(x^*) + \sum_{i=1}^n (h_i^*(y_i^k) + l_i^*(y_i^k)) - \langle x^*, \mathbf{B}^* \mathbf{y}^k \rangle \\ & \quad + \sup_{\mathbf{y} \in D_2} \left(\langle \mathbf{B} x^k, \mathbf{y} \rangle - \sum_{i=1}^n (h_i^*(y_i) + l_i^*(y_i)) \right) \\ & \geq f(x^k) + g(x^k) + \sum_{i=1}^n (h_i \square l_i)(B_i x^k) - \left(f(x^*) + g(x^*) + \sum_{i=1}^n (h_i \square l_i)(B_i x^*) \right). \quad \square \end{aligned}$$

Fix $i \in \{1, \dots, n\}$. The bounded domain conditions in Proposition 5.3 are related to the Lipschitz continuity of the objective functions. Indeed, if h_i is Lipschitz, it follows that $\text{dom}(h_i^*)$ is bounded [5, Proposition 4.4.6]. In addition, $\text{dom}(h_i^* + l_i^*) = \text{dom}(h_i^*) \cap \text{dom}(l_i^*)$. Thus, if h_i^* has bounded domain, so does $h_i^* + l_i^*$.

The bounded subgradient conditions in Proposition 5.3 are satisfied for $h_i \square l_i$ if the infimal convolution is continuous everywhere and the sequence $(B_i x^j)_{j \in \mathbf{N}}$ is convergent. Indeed, in this case $\partial(h_i \square l_i)$ is locally bounded [2, Proposition 16.14(iii)] and hence, the union $\bigcup_{j \in \mathbf{N}} \partial(h_i \square l_i)(B_i x^j)$ is bounded. See [11, Remark 2.2] and [6] for similar remarks in the context of primal-dual FBF and FBS algorithms.

5.2. Two algorithm classes. In this section, we study the algorithms that arise for different classes of maps $(U_j)_{j \in \mathbf{N}}$ and show how to compute the resolvent and forward-backward operators needed in order to apply the PPA, FBS, PRS, and FBF algorithms just as they appear in Section 2.

We fix the following notation for the rest of this section: Let $\mu_{V_i} > 0$ and let $V_i \in \mathcal{S}_{\mu_{V_i}}(\mathcal{H}_i)$ for $i = 0, \dots, n$. Let $\mu_{W_i} > 0$ and let $W_i \in \mathcal{S}_{\mu_{W_i}}(\mathcal{H}_i)$ for $i = 1, \dots, n$. These strongly monotone maps induce metrics on the spaces \mathcal{H}_i for $i = 0, \dots, n$. They can be as simple as “diagonal” metrics, but they can also incorporate second order information. A discussion on the best metric choice is beyond the scope of this paper, so we just refer the reader to [40] for some applications of fixed “diagonal” metrics, and [27] for varying “diagonal” metrics that satisfy conditions akin to Assumption 3.

Now define “block-diagonal” maps

$$\mathbf{V} := V_1 \oplus \dots \oplus V_n \in \mathcal{S}_{\mu_{\mathbf{V}}} \left(\prod_{i=1}^n \mathcal{H}_i \right) \quad \text{and} \quad \mathbf{W} := W_1 \oplus \dots \oplus W_n \in \mathcal{S}_{\mu_{\mathbf{W}}} \left(\prod_{i=1}^n \mathcal{H}_i \right) \quad (5.11)$$

where $\mu_{\mathbf{V}} = \min\{\mu_{V_1}, \dots, \mu_{V_n}\}$, and $\mu_{\mathbf{W}} = \min\{\mu_{W_1}, \dots, \mu_{W_n}\}$. The rest of this section will build three types of metrics from $V_0, \mathbf{V}, \mathbf{W}$.

Finally, note that Part 1 of Proposition 1.2 shows the following: for all $\mathbf{z} \in \mathbf{H}$,

$$\mathbf{z}^+ = J_{U^{-1}(\partial \mathbf{f} + \mathbf{S})}(\mathbf{z}) \quad \Longleftrightarrow \quad U(\mathbf{z} - \mathbf{z}^+) \in \partial \mathbf{f}(\mathbf{z}^+) + \mathbf{S}\mathbf{z}^+. \quad (5.12)$$

See Proposition 5.5, 5.7, and 5.8 for examples of resolvent computations.

5.2.1. First metric class. In this section, our metrics depend on a parameter w , which appears in Algorithm 4. We only use the metric for the case that $w \in \{0, 1/2, 1\}$, but we state all of our results for the general case $w \in \mathbf{R}$. The case $w = 1/2$ first appeared in [9, Theorem 2.1] (for certain \mathbf{V} and V_0), and the case $w = 1$ first appeared in [28, Equation (2.5)] (for certain \mathbf{V} and V_0). See also [46, Relation (3.14)].

PROPOSITION 5.4. *Let $w \in \mathbf{R}$. Assume the setting of Proposition 5.1. Define a map $U_w : \mathbf{H} \rightarrow \mathbf{H}$ as follows: for all $\mathbf{x} = (x, \mathbf{y}) \in \mathbf{H}$,*

$$U_w \mathbf{x} := (V_0 x - w \mathbf{B}^* \mathbf{y}, -w \mathbf{B} x + \mathbf{V} \mathbf{y}). \quad (5.13)$$

Suppose that $w^2 \|\mathbf{V}^{-1/2} \mathbf{B} V_0^{-1/2}\|^2 < 1$. Then U_w is self adjoint and strongly monotone: for all $\mathbf{x} \in \mathbf{H}$,

$$\langle \mathbf{x}, U_w \mathbf{x} \rangle \geq \frac{1}{2} \left(1 - w^2 \|\mathbf{V}^{-1/2} \mathbf{B} V_0^{-1/2}\|^2 \right) \min\{\mu_{V_0}, \mu_{\mathbf{V}}\} (\|x\|^2 + \|\mathbf{y}\|^2). \quad (5.14)$$

Assume the setting of Proposition 5.2. Define a map $U'_w : \mathbf{H} \rightarrow \mathbf{H}$ as follows: for all $\mathbf{x} = (x, \mathbf{v}, \mathbf{y}) \in \mathbf{H}$,

$$U'_w \mathbf{x} := (V_0 x - w \mathbf{B}^* \mathbf{y}, \mathbf{V} \mathbf{y} - w \mathbf{B} x + w \mathbf{v}, w \mathbf{y} + \mathbf{W} \mathbf{v}). \quad (5.15)$$

Suppose that $w^2 \|\mathbf{V}^{-1/2} \mathbf{B} V_0^{-1/2}\|^2 + w^2 \|\mathbf{W}^{-1/2} \mathbf{V}^{-1/2}\|^2 < 1$. Then

$$\begin{aligned} \langle \mathbf{x}, U'_w \mathbf{x} \rangle &\geq \frac{1}{3} \left(1 - w^2 \|\mathbf{V}^{-1/2} \mathbf{B} V_0^{-1/2}\|^2 - w^2 \|\mathbf{W}^{-1/2} \mathbf{V}^{-1/2}\|^2 \right) \\ &\quad \times \min\{\mu_{V_0}, \mu_{\mathbf{V}}, \mu_{\mathbf{W}}\} (\|x\|^2 + \|\mathbf{y}\|^2 + \|\mathbf{v}\|^2). \end{aligned} \quad (5.16)$$

We omit the proof of Proposition 5.4 because Equation (5.14) is shown in [39, Lemma 4.3, Equation (4.14)] when $w = 1$, the extension to general w is straightforward, and Equation (5.16) has nearly the same proof.

Note that our conditions for ergodic convergence in Theorem 3.2 require the metric inducing maps to be almost decreasing up to a summable residual in the Loewner partial ordering \succsim (see Section 1.2). If $w \in \mathbf{R}$ and $((U_w)_j)_{j \in \mathbf{N}}$ is a sequence of maps defined as in Equation (5.13), we have

$$((U_w)_k - (U_w)_{k+1}) \mathbf{x} = ((V_{0,k} - V_{0,k+1}) x, (\mathbf{V}_k - \mathbf{V}_{k+1}) \mathbf{y})$$

for all $\mathbf{x} \in \mathbf{H}$ and $k \in \mathbf{N}$. Thus, if for all $k \in \mathbf{N}$, we have $V_{0,k} \succsim V_{0,k+1}$ and $\mathbf{V}_k \succsim \mathbf{V}_{k+1}$, we can guarantee that the product metric is decreasing (Lemma 1.1). A similar result holds for the level 2 metrics in Equation (5.15).

The following proposition shows how to evaluate the FBS operator under the metrics induced by U_w and U'_w . Note that the results of Proposition 5.5 are not new. The level 1 case with $w \in \{0, 1/2, 1\}$ has appeared implicitly in several papers, including [22, 46, 21]. It has also explicitly appeared in [39, Lemma 4.5]. In addition, the proof of the level 2 case appeared in [9, Equation (2.38)]. Thus, we omit the proof.

PROPOSITION 5.5 (Forward-Backward operators under the first metric class). *Let $w \in \mathbf{R}$. Assume the setting of Propositions 5.1 and 5.4, and suppose that $U_w \in \mathcal{S}_\rho(\mathbf{H})$ (Equation (5.13)) for some $\rho > 0$. Let $\mathbf{z} := (x, \mathbf{y}) \in \mathbf{H}$. Suppose that g, l_1^*, \dots, l_n^* are differentiable. Then $\mathbf{z}^+ := J_{U_w^{-1}(\partial \mathbf{f} + w \mathbf{S})}(\mathbf{z} - U_w^{-1} \nabla \mathbf{g}(\mathbf{z}))$ has the following form: $\mathbf{z}^+ = (x^+, \mathbf{y}^+) \in \mathbf{H}$ where*

$$\begin{aligned} x^+ &= \text{prox}_f^{V_0}(x - V_0^{-1}(w \mathbf{B}^* \mathbf{y} + \nabla g(x))); \\ \text{for } i &= 1, 2, \dots, n, \text{ in parallel do} \\ \quad y_i^+ &= \text{prox}_{h_i^*}^{V_i}(y_i + V_i^{-1}(w B_i(2x^+ - x) - \nabla l_i^*(y_i))); \end{aligned}$$

Assume the setting of Proposition 5.2, and suppose that $U'_w \in \mathcal{S}_\rho(\mathbf{H})$ (Equation (5.15)) for some $\rho > 0$. Let $\mathbf{z} := (x, \mathbf{y}, \mathbf{v}) \in \mathbf{H}$, and suppose that g is differentiable. Then $\mathbf{z}^+ := J_{(U'_w)^{-1}(\partial \mathbf{f} + w \mathbf{S})}(\mathbf{z} - (U'_w)^{-1} \nabla \mathbf{g}(\mathbf{z}))$ has the following form: $\mathbf{z}^+ = (x^+, \mathbf{v}^+, \mathbf{y}^+) \in \mathbf{H}$ where

$$\begin{aligned} x^+ &= \text{prox}_f^{V_0}(x - V_0^{-1}(w \mathbf{B}^* \mathbf{y} + \nabla g(x))); \\ \text{for } i &= 1, 2, \dots, n, \text{ in parallel do} \\ \quad v_i^+ &= \text{prox}_{l_i^*}^{W_i}(v_i + w W_i^{-1} y_i); \\ \quad y_i^+ &= \text{prox}_{h_i^*}^{V_i}(y_i + V_i^{-1}(w B_i(2x^+ - x) - (2v_i^+ - v_i))); \end{aligned}$$

5.3. Second metric class. The following result is similar to [39, Lemma 4.9] (which applies to $(U_w)^{-1}$).

PROPOSITION 5.6. *Assume the setting of Proposition 5.1. Define a map $U_w : \mathbf{H} \rightarrow \mathbf{H}$ as follows: for all $\mathbf{x} = (x, \mathbf{y}) \in \mathbf{H}$,*

$$U_w \mathbf{x} := (V_0 x, (\mathbf{V} - w^2 \mathbf{B} V_0^{-1} \mathbf{B}^*) \mathbf{y}). \quad (5.17)$$

Suppose that $w^2 \|\mathbf{V}^{-1/2} \mathbf{B} V_0^{-1/2}\|^2 < 1$. Then U_w is self adjoint and strongly monotone: for all $\mathbf{x} \in \mathbf{H}$,

$$\langle \mathbf{x}, U_w \mathbf{x} \rangle \geq \min \left\{ \mu_{V_0}, \left(1 - w^2 \|\mathbf{V}^{-1/2} \mathbf{B} V_0^{-1/2}\|^2 \right) \mu_{\mathbf{V}} \right\} (\|x\|^2 + \|\mathbf{y}\|^2). \quad (5.18)$$

Proof. Set $\mathbf{C} = w\mathbf{B}$. For all $\mathbf{y} \in \prod_{i=1}^n \mathcal{H}_i$, we have

$$\begin{aligned} \langle \mathbf{y}, (\mathbf{V} - \mathbf{C} V_0^{-1} \mathbf{C}^*) \mathbf{y} \rangle &= \langle \mathbf{V}^{1/2} \mathbf{y}, \left(I_{\prod_{i=1}^n \mathcal{H}_i} - \mathbf{V}^{-1/2} \mathbf{C} V_0^{-1} \mathbf{C}^* \mathbf{V}^{-1/2} \right) \mathbf{V}^{1/2} \mathbf{y} \rangle \\ &= \langle \mathbf{V} \mathbf{y}, \mathbf{y} \rangle - \langle \mathbf{V}^{1/2} \mathbf{y}, \mathbf{V}^{-1/2} \mathbf{C} V_0^{-1} \mathbf{C}^* \mathbf{V}^{-1/2} \mathbf{V}^{1/2} \mathbf{y} \rangle \\ &\geq \left(1 - \|\mathbf{V}^{-1/2} \mathbf{C}^* V_0^{-1} \mathbf{C} \mathbf{V}^{-1/2}\| \right) \langle \mathbf{V} \mathbf{y}, \mathbf{y} \rangle \\ &\geq \left(1 - w^2 \|\mathbf{V}^{-1/2} \mathbf{B} V_0^{-1/2}\|^2 \right) \mu_{\mathbf{V}} \|\mathbf{y}\|^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \langle \mathbf{x}, U_w \mathbf{x} \rangle &\geq \mu_{V_0} \|x\|^2 + \left(1 - w^2 \|\mathbf{V}^{-1/2} \mathbf{B} V_0^{-1/2}\|^2 \right) \mu_{\mathbf{V}} \|\mathbf{y}\|^2 \\ &\geq \min \left\{ \mu_{V_0}, \left(1 - w^2 \|\mathbf{V}^{-1/2} \mathbf{B} V_0^{-1/2}\|^2 \right) \mu_{\mathbf{V}} \right\} (\|x\|^2 + \|\mathbf{y}\|^2). \quad \square \end{aligned}$$

For simplicity and because it has not yet found an application we do not discuss the generalization of the Equation (5.17) to the level 2 case.

Note that our conditions for ergodic convergence in Theorem 3.2 require the metric inducing maps to be almost decreasing, up to a summable residual, in the Loewner partial ordering \succcurlyeq (see Section 1.2). If $w \in \mathbf{R}$ and $((U_w)_j)_{j \in \mathbf{N}}$ is a sequence of maps defined as in Equation (5.17), we have

$$((U_w)_k - (U_w)_{k+1}) \mathbf{x} = \left((V_{0,k} - V_{0,k+1}) x, \left((\mathbf{V}_k - \mathbf{V}_{k+1}) + w^2 \mathbf{B} (V_{0,k+1}^{-1} - V_{0,k}^{-1}) \mathbf{B}^* \right) \mathbf{y} \right)$$

for all $\mathbf{x} \in \mathbf{H}$ and $k \in \mathbf{N}$. Thus, if for all $k \in \mathbf{N}$, we have $V_{0,k} \succcurlyeq V_{0,k+1}$ and $\mathbf{V}_k \succcurlyeq \mathbf{V}_{k+1}$, the product metric is decreasing (Lemma 1.1).

The following proposition shows how to evaluate the FBS operator under the metric induced by U . Note that Proposition 5.7 appears in [39, Lemma 4.10] for $w = 1$. Thus, we omit the proof.

PROPOSITION 5.7 (Forward-Backward operators under the second metric class). *Assume the setting of Proposition 5.1. Suppose that $f \equiv 0$, and that $U \in \mathcal{S}_\rho(\mathbf{H})$ (Equation (5.17)) for some $\rho > 0$. Let $\mathbf{z} := (x, \mathbf{y}) \in \mathbf{H}$. Suppose that g, l_1^*, \dots, l_n^* are differentiable. Then $\mathbf{z}^+ := J_{U^{-1}(\partial \mathbf{f} + \mathbf{S})}(\mathbf{z} - U^{-1} \nabla \mathbf{g}(\mathbf{z}))$ has the following form: $\mathbf{z}^+ = (x^+, \mathbf{y}^+) \in \mathbf{H}$ where*

$$\begin{aligned} &\text{for } i = 1, 2, \dots, n, \text{ in parallel do} \\ &\quad \lfloor y_i^+ = \text{prox}_{h_i^*}^{V_i} (y_i + V_i^{-1} (w B_i (x - V_0^{-1} (\nabla g(x) + w \mathbf{B}^* \mathbf{y}) - \nabla l_i^*(y_i)))); \\ &\quad x^+ = x - V_0^{-1} (\nabla g(x) + w \mathbf{B}^* \mathbf{y}^+); \end{aligned}$$

Reference	Algorithm	Metric	Level	w	Rates
[15, Algorithm 1]	PPA	(5.13)	1	1	$O(1/(k+1))$ ergodic [15]
[9, Algorithm 2.2]	PPA	(5.15)	2	1	none
[22, 46]	FBS	(5.13)	1	1	$O(1/(k+1))$ ergodic [6]
[16, 18, 39]	FBS	(5.17)	1	1	none
[9, Algorithm 2.1]	PRS	(5.13)	1	1/2	none
[12, Remark 2.9], [35]	PRS	(5.17)	1	0	none
[12, 19]	FBF	(5.17)	1	0	$O(1/(k+1))$ ergodic [11]

TABLE 1

This table lists the original appearance of the algorithms constructed from pairing the metrics in Section 5.2 with the PPA, FBS, PRS, and FBF algorithms applied to Problem 2. See Propositions 5.1 and 5.2 for the definitions of the “level.”

Now consider the special case $w = 0$. In this case, the first and second metric classes agree. The following Proposition with $U = I_{\mathbf{H}}$ appears in [12, Proposition 2.7]. Our generalization is straightforward, so we omit the proof.

PROPOSITION 5.8 (Resolvents of skew operators). *Assume the setting of Proposition 5.1. Let $w \in \mathbf{R}$ and suppose that $U_w \in \mathcal{S}_\rho(\mathbf{H})$ (Equation (5.17)) for some $\rho > 0$. Let $\mathbf{z} := (x, \mathbf{y}) \in \mathbf{H}$. Then $\mathbf{z}^+ := J_{\gamma U^{-1}} \mathbf{S}(\mathbf{z})$ has the following form: $\mathbf{z}^+ = (x^+, \mathbf{y}^+) \in \mathbf{H}$ where*

$$\begin{aligned} x^+ &:= (I_{\mathcal{H}_0} + \gamma^2 V_0 \mathbf{B}^* \mathbf{V} \mathbf{B})^{-1} (x - \gamma V_0 \mathbf{B}^*) \mathbf{y} \\ \mathbf{y}^+ &:= (I_{\prod_{i=1}^n \mathcal{H}_i} + \gamma^2 \mathbf{V} \mathbf{B} V_0 \mathbf{B}^*)^{-1} (\mathbf{y} + \gamma \mathbf{V} \mathbf{B} x) \end{aligned}$$

Generalizing the resolvent operator computation in Proposition 5.8 to the level 2 case is straightforward, though slightly messy. It has not found application in the literature yet, so we omit the statement.

5.4. New and old convergence rates. Table 5.4 lists the application of PPA, FBS, PRS, and FBF algorithms under the metrics introduced in Section 5.2 and indicates which convergence rates have been shown in the literature. We note that, to the best of our knowledge, for all of the methods we discuss, the nonergodic fixed metric convergence rates, the ergodic convergence rates under variable metrics, and the nonergodic/ergodic convergence rates with nonconstant relaxation have never appeared in the literature.

Any pairing between metrics, algorithms, and splittings that does not appear in Table 5.4 is an algorithm where, to the best of our knowledge, no convergence rate has appeared in the literature.

6. Conclusion. In this paper, we provided a convergence rate analysis of a general monotone inclusion problem under the application of four different algorithms. We provided *ergodic* convergence rates under variable metrics, stepsizes, and relaxation, and recovered several known rates in the process. In addition, for three of the algorithms we provided the first *nonergodic* primal-dual gap convergence rates that have appeared in the literature. Finally, we showed how our results imply convergence rates of a large class of primal-dual splitting algorithms. The techniques developed

in this paper are not limited to the four algorithms we chose to study, and the proofs of this paper can be used as a template for proving convergence rates of other special cases of the unifying scheme.

Acknowledgement. We thank Professor Wotao Yin and the two anonymous referees; their comments were invaluable.

REFERENCES

- [1] J.-B. BAILLON AND G. HADDAD, *Quelques propriétés des opérateurs angle-bornés et n -cycliquement monotones*, Israel Journal of Mathematics, 26 (1977), pp. 137–150.
- [2] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011.
- [3] S. BECKER AND P. L. COMBETTES, *An Algorithm for Splitting Parallel Sums of Linearly Composed Monotone Operators, with Applications to Signal Recovery*, arXiv preprint arXiv:1305.5828v1, (2013).
- [4] D. P. BERTSEKAS, *Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey*, Tech. Report LIDS-P-2848, MIT, 2010.
- [5] J. M. BORWEIN AND J. D. VANDERWERFF, *Convex Functions: Constructions, Characterizations and Counterexamples*, vol. 109, Cambridge University Press, 2010.
- [6] R. I. BOŢ AND E. R. CSETNEK, *On the convergence rate of a forward-backward type primal-dual splitting algorithm for convex optimization problems*, Optimization, 64 (2015), pp. 5–23.
- [7] R. I. BOŢ, E. R. CSETNEK, A. HEINRICH, AND C. HENDRICH, *On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems*, Mathematical Programming, 150 (2015), pp. 251–279.
- [8] R. I. BOŢ, E. R. CSETNEK, AND E. NAGY, *Solving systems of monotone inclusions via primal-dual splitting techniques*, Taiwanese Journal of Mathematics, 17 (2013), pp. 1983–2009.
- [9] R. I. BOŢ AND C. HENDRICH, *A Douglas–Rachford Type Primal-Dual Method for Solving Inclusions with Mixtures of Composite and Parallel-Sum Type Monotone Operators*, SIAM Journal on Optimization, 23 (2013), pp. 2541–2565.
- [10] ———, *Solving monotone inclusions involving parallel sums of linearly composed maximally monotone operators*, arXiv preprint arXiv:1306.3191v2, (2013).
- [11] ———, *Convergence Analysis for a Primal-Dual Monotone + Skew Splitting Algorithm with Applications to Total Variation Minimization*, Journal of Mathematical Imaging and Vision, 49 (2014), pp. 551–568.
- [12] L. M. BRICEÑO-ARIAS AND P. L. COMBETTES, *A Monotone+Skew Splitting Model for Composite Monotone Inclusions in Duality*, SIAM Journal on Optimization, 21 (2011), pp. 1230–1250.
- [13] R. E. BRUCK JR., *On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space*, Journal of Mathematical Analysis and Applications, 61 (1977), pp. 159–164.
- [14] A. CHAMBOLLE AND P.-L. LIONS, *Image recovery via total variation minimization and related problems*, Numerische Mathematik, 76 (1997), pp. 167–188.
- [15] A. CHAMBOLLE AND T. POCK, *A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging*, Journal of Mathematical Imaging and Vision, 40 (2011), pp. 120–145.
- [16] P. CHEN, J. HUANG, AND X. ZHANG, *A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration*, Inverse Problems, 29 (2013), pp. 25011–25043.
- [17] P. L. COMBETTES, *Systems of Structured Monotone Inclusions: Duality, Algorithms, and Applications*, SIAM Journal on Optimization, 23 (2013), pp. 2420–2447.
- [18] P. L. COMBETTES, L. CONDAT, J.-C. PESQUET, AND B. C. VŨ, *A forward-backward view of some primal-dual optimization methods in image recovery*, arXiv preprint arXiv:1406.5439v1, (2014).
- [19] P. L. COMBETTES AND J.-C. PESQUET, *Primal-Dual Splitting Algorithm for Solving Inclusions with Mixtures of Composite, Lipschitzian, and Parallel-Sum Type Monotone Operators*, Set-Valued and Variational Analysis, 20 (2012), pp. 307–330.
- [20] P. L. COMBETTES AND B. C. VŨ, *Variable Metric Quasi-Fejér Monotonicity*, Nonlinear Analysis: Theory, Methods & Applications, 78 (2013), pp. 17–31.
- [21] P. L. COMBETTES AND B. C. VŨ, *Variable Metric Forward-Backward Splitting with Applications to Monotone Inclusions in Duality*, Optimization, 63 (2014), pp. 1289–1318.

- [22] L. CONDAT, *A Primal–Dual Splitting Method for Convex Optimization Involving Lipschitzian, Proxiable and Linear Composite Terms*, Journal of Optimization Theory and Applications, 158 (2013), pp. 460–479.
- [23] E. CORMAN AND X. YUAN, *A Generalized Proximal Point Algorithm and Its Convergence Rate*, SIAM Journal on Optimization, 24 (2014), pp. 1614–1638.
- [24] D. DAVIS AND W. YIN, *Convergence rate analysis of several splitting schemes*, arXiv preprint arXiv:1406.4834v3, (2015).
- [25] E. ESSER, X. ZHANG, AND T. F. CHAN, *A General Framework for a Class of First Order Primal-Dual Algorithms for Convex Optimization in Imaging Science*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 1015–1046.
- [26] R. GLOWINSKI AND A. MARROCCO, *Sur l’approximation, par éléments finis d’ordre un, et la résolution, par pénalisation-dualité d’une classe de problèmes de Dirichlet nonlinéaires*, Rev. Française d’Aut. Inf. Rech. Oper. R-2 (1975), pp. 41–76.
- [27] T. GOLDSTEIN, E. ESSER, AND R. BARANIUK, *Adaptive Primal-Dual Hybrid Gradient Methods for Saddle-Point Problems*, arXiv preprint arXiv:1305.0546v2, (2013).
- [28] B. HE AND X. YUAN, *Convergence Analysis of Primal-Dual Algorithms for a Saddle-Point Problem: From Contraction Perspective*, SIAM Journal on Imaging Sciences, 5 (2012), pp. 119–149.
- [29] R. JENATTON, J. MAIRAL, G. OBOZINSKI, AND F. BACH, *Proximal Methods for Hierarchical Sparse Coding*, The Journal of Machine Learning Research, 12 (2011), pp. 2297–2334.
- [30] T. KATO, *Perturbation Theory for Linear Operators*, vol. 132, Springer, 1995.
- [31] N. KOMODAKIS AND J.-C. PESQUET, *Playing with Duality: An Overview of Recent Primal-Dual Approaches for Solving Large-Scale Optimization Problems*, arXiv preprint arXiv:1406.5429v2, (2014).
- [32] S. M. LAVALLE, *Planning Algorithms*, Cambridge University Press, 2006.
- [33] P.-L. LIONS AND B. MERCIER, *Splitting Algorithms for the Sum of Two Nonlinear Operators*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.
- [34] Y. NESTEROV, *Smooth minimization of non-smooth functions*, Mathematical Programming, 103 (2005), pp. 127–152.
- [35] D. O’CONNOR AND L. VANDENBERGHE, *Primal-Dual Decomposition by Operator Splitting and Applications to Image Deblurring*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 1724–1754.
- [36] N. OGURA AND I. YAMADA, *Non-strictly convex minimization over the fixed point set of an asymptotically shrinking nonexpansive mapping*, Numerical Functional Analysis and Optimization, 23 (2002), pp. 113–137.
- [37] L. A. PARENTE, P. A. LOTITO, AND M. V. SOLODOV, *A Class of Inexact Variable Metric Proximal Point Algorithms*, SIAM Journal on Optimization, 19 (2008), pp. 240–260.
- [38] G. B. PASSTY, *Ergodic convergence to a zero of the sum of monotone operators in hilbert space*, Journal of Mathematical Analysis and Applications, 72 (1979), pp. 383–390.
- [39] J.-C. PESQUET AND A. REPETTI, *A Class of Randomized Primal-Dual Algorithms for Distributed Optimization*, arXiv preprint arXiv:1406.6404v3, (2014).
- [40] T. POCK AND A. CHAMBOLLE, *Diagonal preconditioning for first order primal-dual algorithms in convex optimization*, in Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 1762–1769.
- [41] T. POCK, D. CREMERS, H. BISCHOF, AND A. CHAMBOLLE, *An Algorithm for Minimizing the Mumford-Shah Functional*, in Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1133–1140.
- [42] R. T. ROCKAFELLAR, *Monotone Operators and the Proximal Point Algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.
- [43] R. SHEFI AND M. TEOULLE, *Rate of Convergence Analysis of Decomposition Methods Based on the Proximal Method of Multipliers for Convex Minimization*, SIAM Journal on Optimization, 24 (2014), pp. 269–297.
- [44] P. TSENG, *A Modified Forward-Backward Splitting Method for Maximal Monotone Mappings*, SIAM Journal on Control and Optimization, 38 (2000), pp. 431–446.
- [45] B. C. VŨ, *A Variable Metric Extension of the Forward-Backward-Forward Algorithm for Monotone Operators*, Numerical Functional Analysis and Optimization, 34 (2013), pp. 1050–1065.
- [46] ———, *A splitting algorithm for dual monotone inclusions involving cocoercive operators*, Advances in Computational Mathematics, 38 (2013), pp. 667–681.